UNIVERSIDADE FEDERAL DE PERNAMBUCO

Universidade Federal de Pernambuco
Centro de Tecnologia e Geociências
Departamento de Eletrônica e Sistemas

Graduação em Engenharia Eletrônica

Marcelo Herculino Queiroz

Binarização de Documentos Históricos baseado na Estatística

Recife
2024

Marcelo Herculino Queiroz

Binarização de Documentos Históricos baseado na Estatística

Trabalho de Conclusão apresentado ao Curso de Graduação em Engenharia Eletrônica, do Departamento de Eletrônica e Sistemas, da Universidade Federal de Pernambuco, como requisito parcial para obtenção do grau de Bacharel em Engenharia Eletrônica.

Orientador(a): Prof. Marcos Antônio Martins de Almeida, D.Sc.

Recife
2024

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

Queiroz, Marcelo Herculino.
Binarização de Documentos Históricos baseado na Estatística / Marcelo Herculino Queiroz. - Recife, 2024.
97p : il., tab.

Orientador(a): Marcos Antônio Martins de Almeida
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro de Tecnologia e Geociências, Engenharia Eletrônica - Bacharelado, 2024.
Inclui referências, anexos.

1. Informações. 2. Armazenamento. 3. Binarização. 4. Estatística. 5. threshold. I. Almeida, Marcos Antônio Martins de. (Orientação). II. Título.

620 CDD (22.ed.)

Marcelo Herculino Queiroz

Binarização de Documentos Históricos baseado na Estatística

Trabalho de Conclusão apresentado ao Curso de Graduação em Engenharia Eletrônica, do Departamento de Eletrônica e Sistemas, da Universidade Federal de Pernambuco, como requisito parcial para obtenção do grau de Bacharel em Engenharia Eletrônica.

Aprovado em: 16/10/2024

Banca Examinadora

Prof. Marcos Antônio Martins de Almeida, D.Sc.
Universidade Federal de Pernambuco

Prof. Patrícia Silva Lessa, D.Sc.
Universidade Federal de Pernambuco

Prof. Elda Lizandra Fernandes da Silva, M.Sc.
Universidade Federal de Pernambuco

Dedico este trabalho a minha mãe, à senhora Iracema Herculino da Silva cuja presença iluminou cada etapa da minha jornada. Sua constante motivação, carinho e apoio incondicional me deram forças para seguir em frente, mesmo nos momentos mais desafiadores. A você minha mãe, que sempre acreditou no meu potencial, mesmo quando eu duvidei, minha eterna gratidão. Este trabalho é também fruto do seu amor e dedicação, que me impulsionaram a chegar até aqui.

Agradecimentos

Eu gostaria de agradecer a todos que me ajudaram durante esta jornada:

Minha noiva, Adria Lins, que sempre me apoiou e me incentivou a continuar na luta.
Professores: Marcos Martins, Marcelo Menezes e Alexandre Beltrão por toda mentoria e aconselhamento durante todos esses anos.
A minha amiga Willy que sempre me ajudou e me aconselhou.

Aos amigos: Daniel, Ricardo, Gabriel, Rubens, Wallace...

A UFPE.

Finalmente, eu gostaria de agradecer à Deus

Agora quando as pessoas perguntarem: "Quando você vai acabar o curso?", finalmente eu vou poder responder.
ACABEI!

O poder nasce do querer. Sempre que o homem aplicar a veemência e perseverante energia de sua alma a um fim, vencerá os obstáculos, e, se não atingir o alvo fará, pelo menos, coisas admiráveis.

José de Alencar

Resumo do Trabalho de Conclusão de Curso apresentado ao Departamento de Eletrônica e Sistemas, como parte dos requisitos necessários para a obtenção do grau de Bacharel em Engenharia Eletrônica(Eng.)

Binarização de Documentos Históricos baseado na Estatística

Marcelo Herculino Queiroz

A perda de informações ao longo da história é um fenômeno de grande relevância, pois é através do conhecimento que muitas culturas e tradições se perpetuam. Nesse sentido, surge a necessidade do armazenamento de documentos históricos. Atualmente, técnicas de binarização de imagens têm sido empregadas para extrair informações de textos históricos. Assim, este trabalho propõe a união da estatística ao processo de binarização, visando aprimorar a binarização, ao incorporar um *threshold* estatístico. O *threshold* é um valor que separa os *pixels* (menor unidade de uma imagem digital) de uma imagem em categorias. O objetivo é extrair dados de documentos históricos para realizar a análise da significância dos dados através do teste de normalidade das distribuições obtidas. A verificação da normalidade é fundamental para a validade de várias técnicas estatísticas, em especial os métodos paramétricos (técnicas estatísticas que só podem ser realizadas quando os dados seguem a normalidade), como a análise de variância (análise que compara as médias de três ou mais grupos para determinar se há diferenças significativas entre os grupos), que pressupõem que os dados sigam uma distribuição normal (distribuição simétrica em torno da média). A conformidade com este pressuposto é um requisito crucial para a validade dos resultados obtidos. Para garantir a normalidade ou pelo menos aproximar os dados dessa condição, faz-se uso de uma técnica estatística conhecida como *bootstrapping*. Este método cria múltiplas amostras a partir dos dados originais, possibilitando uma análise mais robusta da distribuição dos dados e a realização de ajustes necessários para garantir que atendam aos pressupostos de normalidade.

Ao finalizar o processo de extração e análise dos dados são calculadas as médias locais das amostras e a média global, com o intuito de determinar o *threshold* para a binarização dos documentos históricos, que é o modelo proposto neste trabalho. Além do modelo proposto, são aplicadas outras técnicas de binarização clássicas, que visam realizar uma comparação entre os métodos tradicionais e o modelo desenvolvido. Essa comparação é efetuada utilizando métricas como a relação Sinal-Ruído de Pico (métrica que mede a qualidade de uma imagem) e o mapeamento de *pixels* nas imagens binarizadas, permitindo avaliar a qualidade das binarizações e identificar a técnica mais eficaz para a preservação da integridade das informações contidas nos documentos históricos.

Palavras-chave: Informações; Armazenamento; Binarização; Estatística; *threshold*.

Abstract of Course Conclusion Work, presented to Departament of Eletronic and Systems, as a partial fulfillment of the requirements for the degree of Bachelor of Electronic Engineering(Eng.)

Binarization of Historical Documents based on Statistics

Marcelo Herculino Queiroz

The loss of information throughout history is a highly relevant phenomenon, since it is through knowledge that many cultures and traditions are perpetuated. In this sense, the need for storing historical documents arises. Currently, image binarization techniques have been used to extract information from historical texts. Thus, this work proposes the union of statistics with the binarization process, aiming to improve binarization by incorporating a statistical threshold. The threshold is a value that separates the pixels (the smallest unit of a digital image) of an image into categories. The objective is to extract data from historical documents to perform the analysis of the significance of the data through the normality test of the distributions obtained. Checking for normality is essential for the validity of several statistical techniques, especially parametric methods (statistical techniques that can only be performed when the data follow normality), such as analysis of variance (analysis that compares the means of three or more groups to determine whether there are significant differences between the groups), which assume that the data follow a normal distribution (symmetrical distribution around the mean). Compliance with this assumption is a crucial requirement for the validity of the results obtained. To ensure normality or at least bring the data closer to this condition, a statistical technique known as bootstrapping is used. This method creates multiple samples from the original data, enabling a more robust analysis of the data distribution and the necessary adjustments to ensure that they meet the assumptions of normality.

At the end of the data extraction and analysis process, the local averages of the samples and the global average are calculated in order to determine the threshold for

binarization of historical documents, which is the model proposed in this work. In addition to the proposed model, other classical binarization techniques are applied, which aim to perform a comparison between traditional methods and the developed model. This comparison is made using metrics such as the Peak Signal-to-Noise ratio (a metric that measures the quality of an image) and pixel mapping in the binarized images, allowing the evaluation of the quality of the binarizations and the identification of the most effective technique for preserving the integrity of the information contained in historical documents.

Keywords: Information; Storage; Binarization; Statistics; *threshold*

Lista de Ilustrações

2.1 Vizinhança de um *Pixel* 'p'. 27
2.2 Exemplo de um Histograma 27
2.3 Imagens com suas respectivas variâncias e desvios padrões. 28
2.4 Binarização de uma imagem monocromática utilizando *threshold* T . 30
2.5 Exemplo de Binarização Otsu 33
2.6 Exemplo da Binarização Niblack 35
2.7 Exemplo da Binarização Sauvola 35
2.8 Exemplo de uma Distribuição Gaussiana 40

3.1 Diagrama de blocos do processo de binarização 51
3.2 Amostras da base de dados 52
3.3 Amostragem feita na escrita e no papel 53
3.4 **Quadro 1 -** Algoritmo da Importação do dataset 54
3.5 **Quadro 2 -** Algoritmo do Teste de Anderson Darling 55
3.6 **Quadro 3 -** Algoritmo do *Bootstrapping* 56
3.7 **Quadro 4 -** Algoritmo da ANOVA 57

4.1 Histograma da distribuição dos dados da Escrita e Papel 61
4.2 QQ-Plot da distribuição dos dados da Escrita e do Papel. 62
4.3 Histograma dos dados referente a escrita, antes e depois do *Bootstrapping* 64
4.4 Histograma dos dados referente ao Papel, antes e depois do *Bootstrapping* 64

4.5 Histograma dos dados referente ao Papel e a Escrita, antes do *Bootstrapping*, com o *threshold* 67
4.6 Binarização baseado no modelo proposto 67
4.7 Histograma da Imagem NAB01, antes e depois do processo de binarização. 68
4.8 Resultado da análise de sensibilidade da binarização global aplicada ao modelo proposto, com variações negativas do *threshold* 69
4.9 Resultado da análise de sensibilidade da binarização global aplicada ao modelo proposto, com variações positivas do *threshold* 70
4.10 Comparação das binarizações do modelo Proposto e o modelo automático de Otsu 71
4.11 Resultado das binarizações pelos modelos *Niblack* e Sauvola 72
4.12 Resultado das binarizações Adaptativas pelos modelos da Média e da Gaussiana 72
4.13 Amostras Originais utilizadas para avaliação analítica 73
4.14 Aplicação dos métodos estudados à imagem NAB01 74
4.15 Aplicação dos métodos estudados à imagem PRES01 75
4.16 Aplicação dos métodos estudados à imagem MCA01 75
4.17 Aplicação dos métodos estudados à imagem EBX01 76
4.18 Aplicação dos métodos estudados à imagem ESC01 76
4.19 Aplicação dos métodos estudados à imagem NAB02 77
4.20 Comparação do PSNR da imagem original com os modelos de Binarização. 78
4.21 Normalização da comparação do PSNR da imagem original com modelos de Binarização. 79

3.1 **Quadro 5 -** Algoritmo de Binarização pelo Método Proposto 92
3.2 **Quadro 6 -** Algoritmo da Binarização pela Gaussiana 93
3.3 **Quadro 7 -** Algoritmo de Binarização Otsu 93
3.4 **Quadro 8 -** Algoritmo da Binarização pela Média 94

3.5 **Quadro 9 -** Algoritmo de Binarização *Niblack* 95
3.6 **Quadro 10 -** Algoritmo de Binarização Sauvola 96

Lista de Tabelas

4.1 Resultado do Teste de Normalidade - Anderson Darling. 62

4.2 Resultado do Teste de Normalidade após o *bootstrapping*. 64

4.3 Resultado da Análise de Variância - ANOVA. 66

4.4 Resultado da relação PSNR comparado com a imagem original 78

4.5 Comparação da Normalização dos Resultados do PSNR da imagem original com modelos de Binarização. 79

4.6 Mapeamento de Pontos Semelhantes nas Imagens Binarizadas em Comparação ao Modelo Proposto. 80

4.7 Comparação Percentual do Nível de Semelhança com Base no Modelo Proposto. 80

Lista de Símbolos

σ^2	Variância.
μ	Média.
θ	Ângulo de direção.
T	*Threshold*

Lista de Abreviações

MaZda Macierz Zdarzen.

Ad. Média Adaptativo pela Média.

Ad. Gaussiana Adaptativo pela Gaussiana.

OCR Reconhecimento Óptico de Caracteres.

ROI Regions Of Interest (Regiões de Interesse).

IA Inteligência Artificial.

PSNR Peak Signal-to-Noise Ratio (relação sinal-ruído de pico).

PDI Processamento Digital de Imagens.

UFPE Universidade Federal de Pernambuco.

NAB01 Imagem com alta interferência de ruído na frente e verso do documento.

PRES01 Imagem com nível intermediário de envelhecimento do papel.

MCA01 Imagem com baixa interferência de ruído em seus caracteres.

EBX01 Imagem com alto nível de envelhecimento do papel.

ESC01 Imagem com alto nível de envelhecimento e baixa interferência de ruído.

NAB02 Imagem com alto nível de envelhecimento e alta interferência de ruído.

MCW01 Imagem com alto nível de interferência de ruído e baixo envelhecimento do papel.

NAB03 Imagem com alto nível de interferência de ruído e envelhecimento do papel.

MCA01 Imagem com baixo baixo nível de envelhecimento do papel.

Sumário

1 INTRODUÇÃO **20**

1.1 OBJETIVOS 21

1.1.1 Objetivo Geral 21

1.1.2 Objetivos Específicos 21

1.2 JUSTIFICATIVAS 22

1.3 ESTRUTURA DO TRABALHO 23

2 REFERENCIAL TEÓRICO **25**

2.1 PROCESSAMENTO DIGITAL DE IMAGENS 25

2.2 IMAGEM DIGITAL 26

2.3 HISTOGRAMA 28

2.4 BINARIZAÇÃO 29

2.4.1 **Binarização Global** 30

2.4.2 **Binarização Adaptativa** 33

2.5 FILTRAGEM POR MASCARAMENTO 36

2.6 MATRIZ DE CO-OCORRÊNCIA 36

2.7 ANÁLISE ESTATÍSTICA 37

2.7.1 **Teste de Hipóteses** 38

2.7.2 **Teste de Normalidade** 39

2.7.3 ***Bootstrapping*** 43

2.7.4 **Análise de Variância - ANOVA** 44

2.8 MÉTODOS DE AVALIAÇÃO 45

2.8.1 **Razão Sinal - Ruído de Pico (PSNR)** 45
2.9 *SOFTWARE* E AMBIENTES UTILIZADOS 47
2.9.1 ***Software* MaZda®** 47
2.9.2 ***Software* Excel®** 48
2.9.3 ***Software Python*®** 48

3 METODOLOGIA **49**
3.1 DIAGRAMA DE BLOCOS DO PROCESSO DE BINARIZAÇÃO . . 50
3.2 AQUISIÇÃO DE IMAGENS 50
3.3 PROCEDIMENTOS PARA A EXTRAÇÃO DE CARACTERÍSTICAS 52
3.4 PROCESSAMENTO DE DADOS 54
3.5 ANÁLISE E AJUSTES 56
3.6 PROCESSAMENTO DE DADOS APÓS ÀS ANÁLISES E AJUSTES 57
3.7 BINARIZAÇÃO 57
3.8 AVALIAÇÃO 58

4 RESULTADOS **60**
4.1 VISUALIZAÇÃO DOS RESULTADOS GRÁFICOS ESTATÍSTICOS 60
4.1.1 **Análise e Ajustes** 62
4.1.2 **Cálculo do *Threshold* Estatístico** 66
4.2 RESULTADOS GRÁFICOS 66
4.2.1 **Aplicação dos Métodos em Outras Amostras** 73
4.3 RESULTADOS ANALÍTICOS 77
4.3.1 **Relação Sinal-Ruído de Pico-PSNR** 77
4.3.2 **Mapeamento de *Pixels* nas Imagens Binarizadas** 80

5 CONSIDERAÇÕES FINAIS **82**
5.1 CONCLUSÃO 82
5.2 TRABALHOS FUTUROS 84

Referências **85**

Anexos **87**

1 Expressões Estatísticas Baseadas No Histograma **87**

2 Expressões Estatísticas Baseadas Na Matriz de Co-ocorrência **89**

3 Algoritmos Que Executam Binarizações Globais e Adaptativas **92**

Capítulo 1

INTRODUÇÃO

O armazenamento de informações representa um dos processos culturais mais significativos desenvolvidos ao longo da história pelas civilizações. Este processo não apenas serve como alicerce para a continuidade de suas origens e tradições, mas também desempenha o papel crucial de preservar o conhecimento adquirido ao longo dos anos. Diversas civilizações encontraram meios diversos para registrar e preservar seus conhecimentos, como por exemplo, as tábuas de escrita, utilizadas pelos fenícios, o papiro, desenvolvido pelos egípcios e o papel criado pelos chineses. Os métodos de armazenamento citados anteriormente foram fundamentais para a criação dos primeiros acervos e para o estabelecimento de grandes bibliotecas. Entre as bibliotecas mais renomadas estão as bibliotecas de Alexandria e do Vaticano, ambas desempenharam papéis cruciais no desenvolvimento do conhecimento e da cultura.

Atualmente, o armazenamento de informações em papel enfrenta o desafio da deterioração do material que guarda esses conhecimentos. No entanto, o avanço tecnológico trouxe consigo um meio para preservar informações, tais como, a digitalização de imagens (processo de converter uma imagem analógica, como uma foto impressa, em um formato digital). Todavia, alguns materiais foram danificados com o passar do tempo, antes do processo de digitalização (processo de converter informações impressas ou manuscritas em formato digital), o que dificulta a leitura e interpretação. Outras informações sofreram o processo de envelhecimento do papel,

além da interferência frente e verso, o que gera ruído no processo de captura do texto. Assim, a binarização, processo de converter uma imagem em tons de cinza para uma imagem com apenas dois níveis de cor, preto e branco, surge como uma técnica para reduzir ruídos na legibilidade do material.

Nesse contexto, a proposta deste trabalho é estabelecer um modelo de binarização de documentos históricos incorporando um *threshold* (valor limite usado para classificar dados em categorias) estatístico como meio de otimizar o processo de binarização, garantindo resultados mais precisos e eficazes, com a minimização de erros e uma adaptação dinâmica. Dessa forma, a pesquisa aborda o processamento digital de imagens com o objetivo de obter um conjunto de dados para análise estatística e, consequentemente, determinar o *threshold* ideal para a binarização de documentos históricos. Isso permitirá a comparação com outros métodos automatizados de binarização, além de comparar os resultados por meio de testes de relação sinal-ruído de pico-PSNR (que é uma métrica utilizada para avaliar a qualidade da imagem), e mapeamento de *pixels* nas imagens.

1.1 OBJETIVOS

1.1.1 Objetivo Geral

Este trabalho tem como objetivo extrair características e analisar a textura (padrão visual de variações na superfície capturada) de documentos históricos de imagens digitalizadas, para obter um *threshold* por uma metodologia estatística. A aplicação desse estudo será direcionada à binarização das imagens.

1.1.2 Objetivos Específicos

- Adquirir um banco de imagens;
- Extrair dados através da textura das imagens;
- Modelar dados para processamento;

- Aplicar testes estatísticos;
- Determinar *threshold*;
- Binarizar imagens;
- Comparar modelos de binarização com os métodos relação sinal-ruído de pico (PSNR);
- Avaliar qual o modelo que mais se aproxima do método proposto (modelo de binarização global, que aplica o *threshold* estatístico.), pelo mapeamento de *pixels*;

1.2 JUSTIFICATIVAS

A binarização é um processo fundamental no processamento digital de imagens, e o processo de binarizar têm como objetivo a conversão das intensidades dos *pixels* (menor elemento de uma imagem digital) de uma imagem em uma forma binária, representada por 0 e 1. Este processo é de extrema importância para simplificar a análise de imagens e facilitar a extração de características importantes. Tradicionalmente, técnicas de binarização utilizam métodos de binarização global (utiliza um único *threshold* para toda a imagem) ou adaptativo (utiliza múltiplos *threshold* para a imagem, pois a divide em vários blocos para a aplicação desses valores) para transformar imagens em preto e branco. No entanto, essas abordagens muitas vezes enfrentam desafios em cenários complexos, como imagens com variações de iluminação, ruído ou contrastes não uniformes.

Este Trabalho de Conclusão de Curso (TCC) explora o tema "Binarização de Documentos Históricos Baseado na Estatística", e a escolha deste tema é justificada por várias razões significativas, como na busca da melhor precisão para binarização das imagens, especialmente em imagens com condições de iluminação variáveis e com ruído, nos desafios dos métodos convencionais, que podem não ser suficientemente robustos para imagens com características complexas. A pesquisa pretende abordar

limitações (imagens com condições de iluminação variáveis e com ruído) e avaliar a eficácia das abordagens estatísticas para melhorar a qualidade da binarização. Este estudo impacta em diversas aplicações, pois a binarização precisa é essencial em campos diversos, tais como, reconhecimento de caracteres ópticos (OCR), análise de imagens médicas e sistemas de inspeção automatizados. Melhorar a técnica de binarização pode ter um impacto significativo na eficiência e na precisão desses sistemas.

Portanto, a realização deste TCC é importante para o avanço no campo da binarização de imagens, por meio da aplicação de métodos estatísticos. A pesquisa contribuirá para a melhoria das técnicas existentes com modelos consagrados de binarização. O trabalho também oferecerá novas soluções para problemas complexos, como por exemplo, tornar a binarização um processo dinâmico com base científica. Isto impactará positivamente a prática e o conhecimento na área de processamento digital de imagens (manipulação de imagens digitais por meio de algoritmos).

1.3 ESTRUTURA DO TRABALHO

O conteúdo deste TCC está dividido em cinco capítulos e três anexos. As referências encontram-se nas páginas finais. A seguir, um resumo dos capítulos seguintes do TCC.

Capítulo 2. REFERENCIAL TEÓRICO: Descreve conceitos e modelos relevantes ao estudo, fornecendo uma base sólida para a análise e interpretação dos resultados no processo de binarização.

Capítulo 3. METODOLOGIA: Descreve e justifica os métodos e procedimentos utilizados para a aquisição, análise e interpretação dos dados, aplicados na obtenção de um *threshold* no processo de binarização.

Capítulo 4. RESULTADOS: Descreve e analisa os procedimentos realizados no trabalho, como a extração de dados de uma imagem, a obtenção do *threshold* e a aplicação dos modelos de binarização e suas comparações.

Capítulo 5. CONSIDERAÇÕES FINAIS: Descreve se a proposta do trabalho foi satisfeita, baseada nos resultados. Além de recomendar possíveis direções para estudos futuros

Anexo A. Descreve expressões estatísticas baseadas no histograma.

Anexo B. Descreve expressões estatísticas baseadas na matriz de co-ocorrência.

Anexo C. Descreve algoritmos que executam binarizações globais e adaptativas.

Capítulo 2

REFERENCIAL TEÓRICO

Este capítulo tem o intuito de apresentar fundamentos teóricos para auxiliar na compreensão do trabalho, tais como, o conceito de técnicas estatísticas, algoritmos e métodos de avaliação utilizados no estudo.

2.1 PROCESSAMENTO DIGITAL DE IMAGENS

A história do processamento digital de imagens tem suas raízes no século XIX, seu propósito é fornecer ferramentas para facilitar a extração e identificação de informações de uma imagem para interpretação. As primeiras aplicações de PDI foram identificadas nas áreas industrial (no controle de qualidade e monitoramento de processos) e médica (no diagnóstico por imagem). A evolução das áreas supracitadas foi impulsionada pelo desenvolvimento de algoritmos e *softwares* sofisticados. Os objetivos do processamento digital de imagens variam de acordo com a aplicação, tais como, a restauração da qualidade da imagem, segmentação (processo de dividir uma imagem em partes menores), extração de atributos e reconhecimento de padrões. Essas aplicações têm uso em diversas áreas, como por exemplo, medicina (no diagnóstico de pacientes), visão computacional (campo da IA, Inteligência Artificial, que ensina computadores a interpretar e entender imagens ou vídeos, com aplicação no reconhecimento de objetos e análise visual) e entretenimento (no processamento de imagens em jogos e filmes). No presente trabalho, serão abordados

os temas de segmentação e extração de características.

2.2 IMAGEM DIGITAL

Para compreender o processamento de imagens, é essencial entender a definição de uma imagem. A imagem é a representação visual de um objeto, cena ou conceito, definida matematicamente como uma função bidimensional $f(x, y)$, onde a amplitude da função representa o nível de intensidade de tons de RGB da imagem, também conhecido como *pixel*. Em termos técnicos, uma imagem é composta por *pixels*, que são as unidades fundamentais de uma imagem digital (representação visual composta por *pixels*). A quantidade de *pixels* por área determina a resolução da imagem, que é uma medida na qual descreve a quantidade de detalhes de uma determinada cena. Uma alta resolução resulta em uma imagem com mais detalhes, enquanto uma baixa resolução resulta em uma imagem borrada (Gonzalez e Woods, 2010).

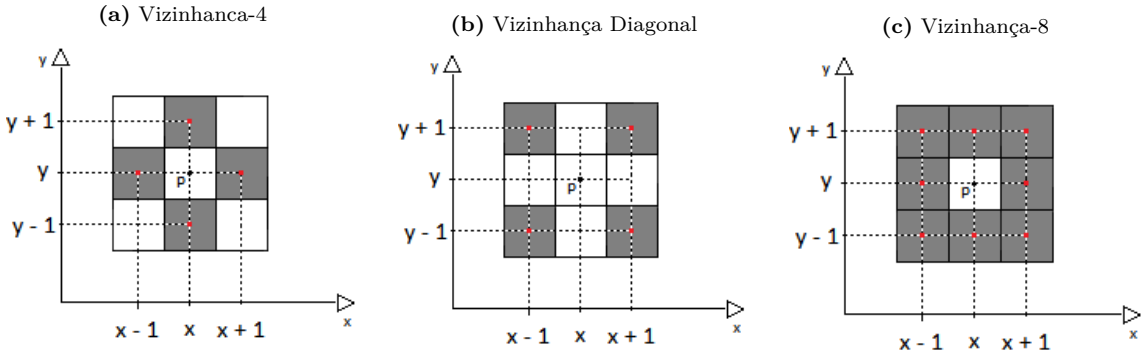
A manipulação individual dos *pixels* e de sua vizinhança é uma prática comum no processamento de imagens. A vizinhança pode ser definida de várias formas, dependendo da aplicação, mas geralmente é compreendida como os *pixels* próximos ao *pixel* analisado. As definições mais comuns são a vizinhança-4, vizinhança diagonal e a vizinhança-8.

A vizinhança-4 de um*pixel* inclui os *pixels* que estão diretamente ligados ao *pixel* em questão na vertical e horizontal, conforme a Figura 2.1(a). A vizinhança diagonal, está relacionado as diagonais do *pixel* de referência, como mostra a Figura 2.1(b). Já a vizinhança-8 de um *pixel*, engloba os *pixels* da vizinhança-4 e os *pixels* das diagonais, mostrado na Figura 2.1(c) e na Equação 2.1. Assim, ao considerar um *pixel* 'p' de coordenadas (x,y), são levados em conta 4 vizinhos nas direções horizontais e verticais, cujas coordenadas são (x + 1, y), (x - 1, y), (x, y + 1) e (x, y - 1). Esses *pixels* formam a chamada "4-vizinhança" de 'p', identificada como $N_4(p)$. Seguindo o mesmo raciocínio para a "8-vizinhança" de 'p', temos a união dos vizinhos nas direções horizontais, verticais e diagonais. Os quatro vizinhos diagonais

de ‘p’ representados pelos *pixels* de coordenadas (x - 1, y - 1), (x - 1, y + 1), (x + 1, y - 1) e (x + 1, y + 1), constituindo o conjunto $N_d(p)$ (Marques Filho e Neto, 1999).

$$N_8(p) = N_4(p) \cup N_d(p) \quad . \tag{2.1}$$

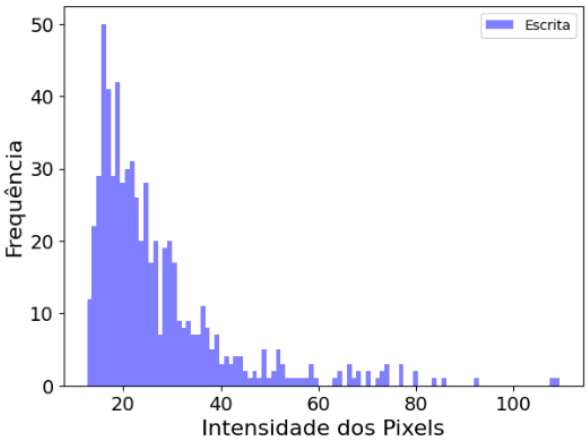
Figura 2.1: Vizinhança de um *Pixel* ‘p’.



(Fonte: Adaptado de MARQUES FILHO, Ogê; VIEIRA NETO, Hugo. *Processamento Digital de Imagens*. Rio de Janeiro: Brasport, 1999, p. 26.)

A importância dos *pixels* e sua vizinhança está na avaliação da intensidade dos níveis de tonalidade de cinza em uma imagem, revelando sua luminosidade e sendo representada graficamente num histograma. Esse tipo de gráfico, oferece uma visão panorâmica da distribuição de frequência dos eventos, como demonstrado na Figura 2.2.

Figura 2.2: Exemplo de um Histograma

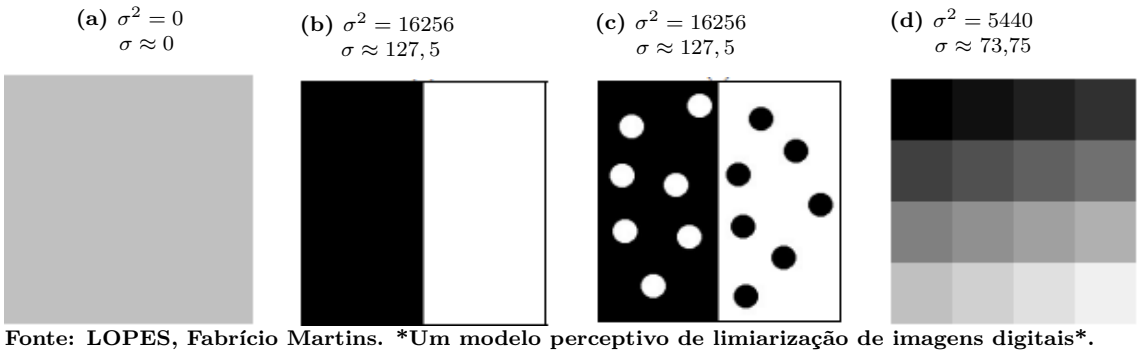


(Fonte: Autor da Figura)

2.3 HISTOGRAMA

O histograma é uma representação gráfica de frequência de dados estatísticos descritivos, divididos em intervalos denominados classes. Essa representação proporciona uma visão geral da distribuição de uma variável, sendo aplicável à análise da intensidade de *pixels* em uma imagem, seja através da intensidade de cores ou dos níveis de cinza. O histograma é uma ferramenta útil para calcular a intensidade média dos *pixels* de uma imagem, representada pelo primeiro momento, e a variância (medida que indica o quanto os valores de um conjunto de dados variam em relação à média) representada pelo segundo momento, os quais desempenham um papel importante na caracterização da iluminação durante a aquisição da imagem. A variância representada por σ^2, em particular, indica a variação das intensidades das tonalidades de cinza em relação à sua média μ, o que caracteriza o contraste da imagem. O histograma pode ser utilizado para realçar, comprimir e segmentar imagens. Dessa forma, é fornecida uma visualização da frequência dos níveis de tonalidade de cinza. Isso é útil para avaliar, monitorar e otimizar o processo de realce usado na manipulação dos dados da imagem (Nunes, 2006).

Figura 2.3: Imagens com suas respectivas variâncias e desvios padrões.



Fonte: LOPES, Fabrício Martins. *Um modelo perceptivo de limiarização de imagens digitais*. Universidade Federal do Paraná, 2003, p. 10.

A Figura 2.3 ilustra imagens digitais e suas respectivas variâncias. A Figura 2.3(a) apresenta variância nula, pois todos os *pixels* têm a mesma intensidade. As Figuras 2.3(b) e 2.3(c) mostram variância máxima, pois seus *pixels* representam a mesma intensidade máxima nos extremos. A Figura 2.3(d) tem variância média entre

seus elementos, com sua intensidade variando de um extremo a outro.Importante ressaltar que a variância não se altera com a disposição espacial da informação na imagem, como ilustrado nas Figuras 2.3(b) e 2.3(c)(Lopes, 2003).

O realce é o processo de melhoria ou destaque de características específicas em uma imagem para torná-la mais perceptível ou detalhada. A compressão é o processo de reduzir o tamanho de uma imagem mantendo ao máximo a qualidade visual, enquanto que a segmentação é o processo de dividir uma imagem em regiões com base em sua cor, intensidade, textura ou forma. Existem várias técnicas para a segmentação, como binarização por *Thresholding* (processo de converter uma imagem em preto e branco, comparando cada *pixel* com um valor limite), segmentação por regiões (agrupa *pixels* semelhantes em uma área contínua, baseando-se em características como cor ou intensidade), bordas (identifica os contornos dos objetos, detectando mudanças bruscas na intensidade ou cor entre diferentes áreas), contorno (separa objetos em uma imagem com base nas bordas ou limites entre eles), segmentação por *Watershed* (trata a imagem como um terreno topográfico, onde os *pixels* com intensidades de cinza mais altas são como picos e os *pixels* com intensidades mais baixas são como vales) e aprendizagem de máquina (campo da IA em que computadores aprendem a fazer previsões ou tomar decisões a partir de dados, sem programação direta). Neste trabalho, especificamente, será utilizada a binarização por *Thresholding*.

2.4 BINARIZAÇÃO

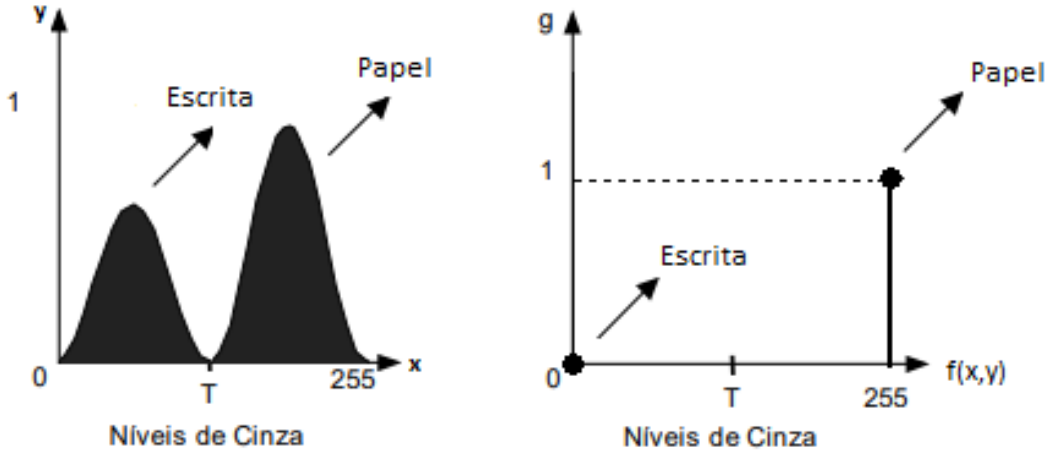
Antes de abordar o conceito de Binarização, é necessário introduzir o conceito de classe, que no contexto do processamento de imagens, refere-se às categorias ou grupos nos quais os *pixels* de uma imagem são classificados, como mostra a Figura 2.4(a).

A binarização é uma técnica baseada em separar regiões de uma imagem em duas classes. Seu processo consiste na bipartição do histograma com base em um certo valor, chamado de *threshold*, usado para binarizar a imagem. Neste processo,

os *pixels* são distribuídos em duas classes: uma que possui intensidades acima do *threshold* e outra classe com os *pixels* que possuem intensidades abaixo dele, como representado nos histogramas da Figura 2.4(b). A interpretação da classificação dos valores dos *pixels* são dependentes do dispositivo, alguns sistemas interpretam o valor 0 como preto e 1 como branco, enquanto outros invertem os sentidos destes valores. Toda essa abordagem é realizada por meio de um processo de mascaramento, que é o processo de aplicar uma máscara para isolar ou ocultar partes específicas de uma imagem, o que permite que apenas as áreas de interesse sejam analisadas, enquanto o restante da imagem é ignorado. (Marques Filho e Neto, 1999).

Figura 2.4: Binarização de uma imagem monocromática utilizando *threshold* T

(a) Histograma da Imagem Original **(b)** Histograma da Imagem Binarizada



Existem vários tipos de binarização que podem ser aplicados em diferentes contextos de processamento de imagens. Alguns dos tipos mais frequentes utilizados são as binarizações global e a Adaptativa, que serão abordados a seguir.

2.4.1 Binarização Global

Na Binarização global, um único valor de *threshold* é empregado para segmentar intensidades de *pixels* de toda a imagem, este método é frequentemente empregado onde o histograma de uma imagem tem intensidades com um padrão de dois picos distintos, representando duas classes distintas de *pixels*.

Matematicamente, a binarização global é descrita na Equação 2.2:

$$g(x,y) = \begin{cases} 0, \text{ escrita (objeto) se } f(x,y) \geq \text{T}. \\ 1, \text{ papel (fundo) se } f(x,y) < \text{T}. \end{cases} \quad . \tag{2.2}$$

Em que $f(x,y)$ é a função que descreve imagem original, e 'T' é o valor do *threshold* que delimita a imagem binarizada em $g(f(x,y))$ como visto na Figura 2.4(b) (Victer e Silva, 2019)(Ribas et al., 2013).

Nas binarizações globais utilizadas neste trabalho aplicou-se dois métodos: o proposto e o método de Otsu (técnica de binarização que determina automaticamente o *threshold*, maximizando a variância entre as classes). No método proposto, foi necessário determinar um *threshold* através de processos estatísticos, que é justamente a proposta abordada, que será vista na metodologia. Existem outros métodos de binarização global, mas não foram utilizados.

Método Otsu

O método de Otsu é uma técnica simples de binarização, baseada no histograma da imagem. Seu objetivo é determinar o valor ideal de um *threshold*, separando os elementos em dois grupos: os que pertencem ao fundo e os que pertencem a frente da imagem. O método Otsu executa todos os valores possíveis de *threshold* na imagem, buscando aquele que minimiza a soma da variância intraclasse (medida da variação entre os elementos dentro de um mesmo grupo ou classe) da imagem. Este cálculo é expresso pela Equação 2.3, apresentada a seguir:

$$\sigma_W^2 = W_b\sigma_b^2 + W_f\sigma_f^2 \quad . \tag{2.3}$$

Em que,

- σ_W^2 é a variância intraclasse, mede a variação dentro das classes (fundo e frente da imagem);
- W_b e W_f são os pesos para cada classe, representando a probabilidade de um *pixel* pertencer à classe de fundo (*background*) ou de frente (*foreground*);

- σ_b^2 e σ_f^2 são as variâncias das intensidades do *pixel* dentro das classes (*background* ou *foreground*).

Inicialmente, essa abordagem calcula a variância intraclasse para todos os possíveis *thresholds*. No entanto, Otsu propôs foi trocar o foco do cálculo direto da variância intraclasse, equação 2.3, citada anteriormente, pela variância interclasse (medida usada para avaliar a separação entre duas classes de pixels) equação 2.4, na de busca maximizar este valor. Assim há uma redução no custo computacional do algoritmo, pois não é necessário processar todos os pixels diretamente. Em vez disso, ele trabalha com o histograma da imagem. Esta troca de processo de trabalho resulta no mesmo *threshold* e, portanto, na binarização ideal. A expressão para a variância interclasse é dada pela Equação 2.4:

$$\sigma_B^2 = \sigma^2 - \sigma_W^2 = W_b(\mu_b - \mu)^2 + W_f(\mu_f - \mu)^2 \quad . \tag{2.4}$$

Em que,

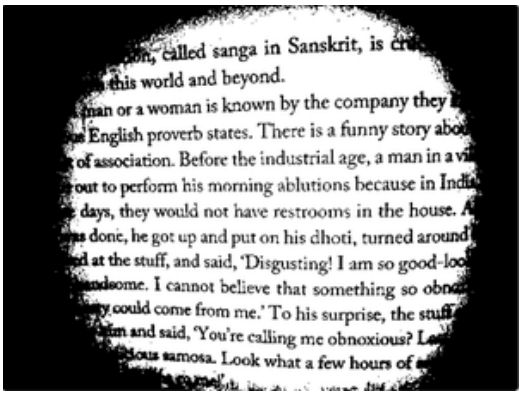
- σ_B^2 é a variância interclasse, mede a variação entre as classes (*background* e *foreground*).
- σ é a variância total da população ou dos dados.
- σ_W^2 é a variância intraclasse, mede a variação dentro das classes.
- W_b e W_f são os pesos para cada classe, representando a probabilidade de um *pixel* pertencer à classe de fundo (*background*) ou de frente (*foreground*).
- μ_b e μ_f média das classes.
- μ é a média total dos dados;

A Equação 2.4 também pode ser escrita, como mostra a equação 2.5

$$\sigma_B^2 = W_b W_f(\mu_b - \mu_f)^2 \quad . \tag{2.5}$$

Essa abordagem torna o método de Otsu não apenas eficiente do ponto de vista computacional, mas também eficaz na binarização de imagens, garantindo resultados de qualidade. Um exemplo do método Otsu pode ser visto na Figura 2.5(Pereira, 2006)(Torok, 2016).

Figura 2.5: Exemplo de Binarização Otsu



(Fonte: Autor da Figura)

2.4.2 Binarização Adaptativa

Na binarização adaptativa, também conhecida como binarização local, são utilizados múltiplos valores de *threshold* com o objetivo de obter melhores resultados. A binarização adaptativa baseia-se em dividir a imagem em regiões menores, chamadas de janelas ou blocos, que definem a área local da imagem para o cálculo do *threshold*. Para cada uma dessas regiões, um *threshold* é calculado. Este valor de *threshold* pode ser determinado de diversas maneiras, tais como, a média (valor que se obtém somando todos os números de um conjunto e dividindo o resultado pela quantidade de números) dos valores de intensidade, a mediana (valor que separa um conjunto de dados em duas partes iguais, sendo o número do meio quando os dados estão ordenados) ou outros métodos estatísticos. O tamanho da janela também deve ser selecionado, pois o resultado também depende deste parâmetro. Janelas maiores garantem uma boa estimativa do valor médio, enquanto janelas menores evitam distorções devido à não uniformidade do fundo. A binarização adaptativa é eficaz para imagens com iluminação não uniforme ou com áreas de intensidades muito diferentes, permitindo uma segmentação mais precisa e robusta (Victer e Silva, 2019).

Os métodos adaptativos utilizados neste trabalho são os de binarização adaptativa pela média, adaptativa pela Gaussiana, *Niblack* e Sauvola, que serão explicados mais adiante. Eles são amplamente utilizados em aplicações como Reconhecimento Óptico de Caracteres (OCR), segmentação de imagens médicas, processamento de documentos e outros.

Binarização Adaptativa pela Média

Binarização adaptativa pela média é um método de processamento local, baseado no cálculo do valor médio dos níveis de intensidade dos *pixels*. Seu processo consiste em dividir a imagem em regiões menores, chamadas janelas, calcular o valor médio para cada uma e binarizar a região correspondente.

Binarização Adaptativa pela Gaussiana

Binarização Adaptativa pela Gaussiana é outra técnica de binarização local usada para segmentar imagens. Baseia-se na distribuição Gaussiana para obter um valor adaptativo que será utilizado para binarizar a região correspondente. Seu processo consiste em dividir a imagem em regiões menores e, para cada região, calcular o valor médio e o desvio padrão dos níveis de intensidade dos *pixels*. Esses valores são então utilizados na distribuição Gaussiana para determinar o *threshold* e binarizar a região correspondente. Esse método leva em consideração as variações locais de intensidade. Assim, para cada *pixel* na imagem, a distribuição Gaussiana é calculada em sua vizinhança, estimando um *threshold* local.

Niblack

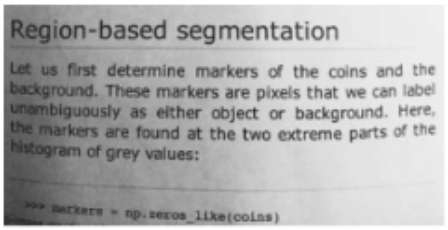
Desenvolvido em 1986 por Wayne Niblack, este método automático de binarização, tem como base calcular os *thresholds* locais, de acordo com a média local 'μ' e o desvio padrão 'σ' das intensidades dos *pixels*, que estão na vizinhança do *pixel* analisado. O método consiste em construir uma superfície limite de valores de cinza, que são computados em uma pequena vizinhança ao redor de cada *pixel*, onde 'k' é uma constante usada na construção da superfície limite dos valores de cinza. Esse

método tende a produzir ruído em imagens com muitos tons de cinza, como mostra a Figura 2.6(b) que é a binarização *Niblack* da Figura original 2.6(a). A fórmula para o método *Niblack* é vista na Equação 2.6 (Seixas et al., 2008).

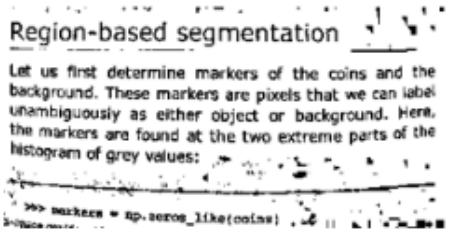
$$T = \mu + k\sigma \quad . \tag{2.6}$$

Figura 2.6: Exemplo da Binarização Niblack

(a) Imagem Original



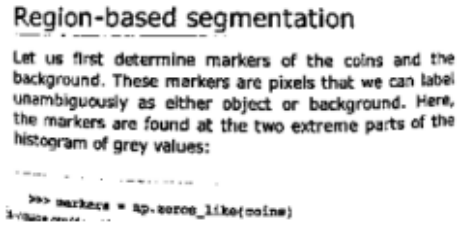
(b) Binarização*Niblack*



Sauvola

O método Sauvola foi desenvolvido em 2000 por J. Sauvola e M. Pietikainen. Considerado uma extensão do método *Niblack*, ele calcula *thresholds* locais com base na média e desvio padrão das intensidades dos *pixels* que estão na vizinhança do *pixel* analisado. Esse método é especialmente adequado para imagens de documentos digitalizados, onde a iluminação pode ser irregular e o contraste pode ser baixo, pois adiciona a hipótese de níveis de tonalidade de cinza, pré-determinados, para os objetos e para o fundo da imagem, como mostra a Figura 2.7(Bertholdo, 2007).

Figura 2.7: Exemplo da Binarização Sauvola



(Fonte: Autor da Figura)

O *threshold* que representa o modelo Sauvola é descrita na Equação 2.7:

$$T = \mu[1 + k(\frac{\sigma}{R} - 1)] \quad . \tag{2.7}$$

A diferença fundamental entre Sauvola e *Niblack* está na maneira de calcular o *threshold*, onde 'k' é um parâmetro que controla a sensibilidade do método, e 'R' é o valor máximo do desvio padrão na vizinhança (Ikeda, 2011).

2.5 FILTRAGEM POR MASCARAMENTO

Um conceito importante para o processamento de imagens é a filtragem por mascaramento. Este processo se aplica na correção do sombreamento de imagens, pois serve para melhorar o processo de segmentação, destaca características importantes e remove ruídos. Neste trabalho, a filtragem supracitada foi utilizado com o objetivo de extrair atributos da textura da imagem. O processo aplicado consiste em selecionar ROI's (Regiões de Interesse), aleatórias na imagem, que identifica o objeto e o respectivo fundo, para extrair dados, que possam identificar dois picos distintos de uma distribuição no histograma. Dessa forma, o processo consiste em multiplicar as regiões de interesse por uma máscara com valor '1' e as demais regiões por '0'. O *software* utilizado nesse procedimento foi o MaZda®, que além de extrair características de textura, fornece informações específicas sobre as relações espaciais entre os níveis de tonalidade de cinza através da matriz de co-ocorrência, que é uma técnica para quantificar as relações entre os *pixels*.

2.6 MATRIZ DE CO-OCORRÊNCIA

A matriz de co-ocorrência se destaca como uma ferramenta essencial no campo de processamento de imagens, sendo utilizada para extrair características texturais da imagem. Trata-se de uma representação estatística que define a distribuição das intensidades de *pixels*. O objetivo da matriz em questão é de classificar a imagem em um conjunto de N classes, com base nos padrões dos níveis de tonalidade de cinza. Nesse contexto, a imagem é dividida em regiões amostradas de forma a permitir que

seus *pixels* possam pertencer, simultaneamente, a múltiplas regiões, melhorando assim a determinação da classe a que cada *pixel* pertence.

Os elementos de uma matriz de co-ocorrência descrevem a frequência com que ocorrem transições nos níveis de cinza em uma imagem. Ou seja, um elemento na linha 'i' e coluna 'j' com valor 'p' (representa a frequência com que uma determinada transição de intensidade ocorre entre dois pixels em uma imagem) nessa matriz. A matriz é definida como um histograma de segunda ordem e utilizada como uma abordagem para a análise de textura (Schwartz e Pedrini, 2003)(Ito et al., 2009).

Em termos simples, a matriz de co-ocorrência $C(i, j, d, \theta)$ calcula a frequência com que os pares de *pixels* aparecem juntos, a uma distância 'd', em uma determinada direção 'θ', da imagem da forma mostrada na Equação 2.8.

$$C_o(i,j,d,\theta) = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} = \begin{cases} 1; \text{ se } f(x,y) = i \text{ e } f(x + dcos(\theta), y + dsin(\theta)) = j \ . \\ \\ 0; \text{ caso contrário.} \end{cases} . \tag{2.8}$$

Em que,

- f(x,y) é o valor de intensidade do *pixel* na posição (x,y) da imagem;
- d: a distância entre os *pixels* para os quais estamos calculando a co-ocorrência
- M e N são as dimensões da imagem, largura e altura, respectivamente;
- Θ é o ângulo de direção (0°, 45°, 90°, 135°) em que a coocorrência é calculada.

Outras equações da matriz coocorrência se encontram no anexo B.

2.7 ANÁLISE ESTATÍSTICA

Este tópico tem como base apresentar os métodos estatísticos utilizados neste estudo, com o objetivo de analisar os dados e inferir na tomada de decisões. Com esse propósito, são apresentados o teste de normalidade e a análise de variância

(ANOVA). O teste de normalidade é uma análise estatística que verifica se a distribuição dos dados é simétrica em torno da sua média, cujo o gráfico têm o formato de sino. A ANOVA é um teste que analisa a variabilidade dos dados (como os valores em um conjunto de dados se dispersam em torno da média) e divide essas informações em diferentes fontes de variação, para determinar se existe uma diferença significativa entre as médias de três ou mais grupos. Em outras palavras, ela analisa como a variância dos dados é distribuída entre diferentes fontes de variação. Portanto, embora o nome "análise de variância"possa sugerir uma técnica focada na variação dos dados em si, o principal objetivo da ANOVA é comparar as médias dos grupos, fazendo isso, ao examinar como a variância entre os grupos se compara à variância dentro dos grupos. Todos esses testes estatísticos são baseados em testes de hipóteses (método estatístico usado para determinar se uma afirmação ou suposição sobre um parâmetro de uma população é verdadeira ou falsa, com base em dados amostrais), os quais envolvem exatamente duas hipóteses sobre a população em estudo.

2.7.1 Teste de Hipóteses

O teste de hipóteses é um método de averiguação sobre a veracidade de uma afirmação, associado a um risco máximo de erro. Em outras palavras, por definição, um teste de hipóteses é uma ferramenta para aceitar ou rejeitar uma hipótese, com base nas informações fornecidas pelos dados coletados em uma amostra.

Devido à maneira como as análises são realizadas, cada teste de hipóteses inclui exatamente duas hipóteses sobre a população em estudo. A primeira é a hipótese nula (H0), considerada verdadeira até que se prove o contrário. A segunda é a hipótese alternativa (H1), que representa uma afirmação contrária aquela definida pela hipótese nula. Essa suposições são feitas para verificar se uma determinada afirmação deve ser aceita ou rejeitada. O teste citado, compara a estatística do teste com o valor crítico ou o p-valor (p-value). Portanto, se a estatística do teste estiver na região crítica ou se o p-valor for menor que α (nível de significância), a hipótese

nula é rejeitada. O valor 'p' considerado no teste é denominado "nível descritivo do teste"ou "probabilidade de significância", que geralmente é estabelecido em 5% (Broch e Ferreira, 2012)(Hirakata et al., 2019).

2.7.2 Teste de Normalidade

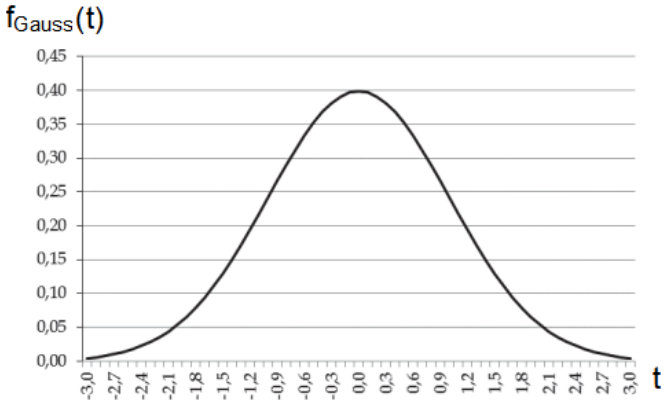
Compreende a configuração da distribuição de probabilidade de uma variável aleatória é crucial e, por vezes, indispensável em contextos estatísticos. Uma vez que a forma da distribuição seja identificada, torna-se viável estimar seus parâmetros, construir intervalos de confiança e realizar testes de hipóteses. Assim, a descrição das distribuições de probabilidade pode assumir várias abordagens, sendo possível caracterizá-las através de diferentes aspectos: pela sua função densidade (probabilidade de uma variável aleatória contínua assumir certos valores), pela sua função característica (função que captura toda a informação sobre sua distribuição de probabilidade), pela sua respectiva função geradora de momentos (função que descreve a distribuição de uma variável aleatória ao gerar seus momentos, como média e variância) e pelo conjunto de seus respectivos momentos (conjunto de medidas estatísticas que descrevem as características de uma distribuição de probabilidade). Na prática, é comum empregar os quatro primeiros momentos para descrever uma distribuição amostral: (a) o primeiro momento fornece uma medida de localização ou tendência central (média, mediana e moda(valor que aparece com mais frequência em um conjunto de dados)); (b) o segundo momento representa uma medida de dispersão (variância, desvio padrão, coeficiente de variação (medida de dispersão que expressa a variabilidade de um conjunto de dados em relação à média) e amplitude, medidas estatísticas que descrevem o quanto os dados em um conjunto variam ou se espalham em relação à média); (c) o terceiro momento oferece uma medida de *skewness*(assimetria), refere-se ao grau em que a distribuição dos dados não é simétrica em torno da média; (d) o quarto momento, conhecido como *kurtosis* (curtose), quantifica a proeminência do pico e a forma das caudas da curva de distribuição (Pino, 2014).

Com o intuito de analisar a distribuição de dados, o passo essencial é realizar o teste de normalidade, que é uma ferramenta estatística utilizada para verificar se os dados de uma amostra seguem a distribuição normal (Gaussiana), distribuição de probabilidade contínua em forma de sino, caracterizada pela simetria em torno da média. Uma variável aleatória 'γ' tem distribuição normal, com média 'μ' e variância 'σ^2', e é representada por $\gamma \sim N(\mu, \sigma^2)$, se sua função densidade de probabilidade é dada pela seguinte Equação 2.9:

$$f_{Gauss}(t) = (2\pi\sigma^2)^{-\frac{1}{2}} exp[-\frac{(t-\mu)^2}{2\sigma^2}] \quad . \tag{2.9}$$

para $-\infty < t < \infty$, e $\sigma > 0$. A distribuição normal padrão têm média zero e variância um: $\gamma \sim N(0, 1)$, e terá o formato exibido na Figura 2.8.

Figura 2.8: Exemplo de uma Distribuição Gaussiana



Distribuição Gaussiana

O teste de normalidade é muito importante, pois valida vários testes estatísticos, tais como, os testes paramétricos, que são técnicas com suposições específicas sobre a distribuição de dados, que são feitas suposições estimando parâmetros. Existem vários tipos de teste de normalidade, sendo os mais conhecidos, o teste de Shapiro-Wilk, o Kolmogorov-Smirnov e o Anderson-Darling.

O Teste de Shapiro-Wilk possui uma abordagem que analisa a relação entre os dados observados e os dados esperados de uma distribuição normal, já os testes Kolmogorov-Smirnov e Anderson-Darling são baseados na função de distribuição

empírica, que consiste em compará-la com a função de distribuição acumulada da normal (Meneses e Almeida, 2012)(Pino, 2014).

Teste de Kolmogorov-Smirnov

O teste de Kolmogorov-Smirnov é uma técnica estatística não paramétrica usada para avaliar se uma amostra segue uma distribuição específica, como a distribuição normal. Este teste compara a distribuição acumulada observada dos dados, com a distribuição acumulada esperada (teórica). Então, o teste calcula a diferença máxima entre as duas respectivas funções acumuladas. O resultado do teste fornece uma estatística D, que é comparada com valores críticos da Tabela de Kolmogorov-Smirnov, para determinar se há evidências suficientes, para rejeitar a hipótese nula de que os dados seguem a distribuição teórica. Na prática, se o valor-p associado ao teste for menor que um nível de significância pré-determinado, rejeita-se a hipótese nula, indicando que os dados não seguem a distribuição teórica proposta (Pino, 2014).

Teste de Anderson Darling

O teste de Anderson-Darling é outra técnica estatística utilizada para avaliar se uma amostra de dados segue uma distribuição normal. Derivado do teste de Kolmogorov-Smirnov, ele concentra-se nas diferenças entre a distribuição acumulada observada e a distribuição acumulada teórica. No entanto, o teste de Anderson-Darling atribui pesos maiores às variações na cauda da distribuição, o que aumenta a sensibilidade do teste a desvios nas nas caudas da distribuição de dados. Por essa razão este teste, foi escolhido para ser implementado neste estudo. O procedimento envolve o cálculo de uma estatística de teste específica, geralmente denotada como A^2 e descrita na Equação 2.10, e sua comparação com os valores críticos da Tabela de Anderson-Darling.

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}[\frac{2i-1}{n}(ln(F(X_{(i)})) + ln(1 - F(X_{(n+1-i)})))] \quad . \tag{2.10}$$

Em que:

- A^2 é a estatística de teste de Anderson-Darling;
- n é o tamanho da amostra;
- $X_{(i)}$ é o i-ésimo menor valor ordenado;
- $F(X_{(i)})$ é a função de distribuição acumulada empírica dos dados nos pontos $X_{(i)}$.

Assim, como em outros testes de aderência, se o valor-p associado ao teste de Anderson-Darling for menor que um nível de significância pré-determinado, a hipótese nula de que os dados seguem a distribuição teórica proposta é rejeitada (Pino, 2014).

Teste de Shapiro-Wilk

O Teste de Shapiro-Wilk é uma técnica baseada na comparação entre a distribuição dos dados amostrais e uma distribuição normal, utilizada para avaliar se uma amostra segue uma distribuição específica, como a distribuição normal. Esse teste utiliza uma razão entre duas estimativas obtidas de estatísticas de ordem (medidas que organizam dados em uma sequência): uma estimativa ponderada de mínimos quadrados, que considera uma população normalmente distribuída, e a estimativa não viesada (é uma maneira de calcular um valor que, em média, acerta o verdadeiro valor de um parâmetro) da variância amostral, para qualquer população. O teste se baseia no fato de que a variável $\gamma \approx N(\mu, \sigma^2)$ pode ser expressa como $y_{sw} = \mu + \sigma x_{sw}$, onde $X \approx N(0, 1)$, y_{sw} (é o valor esperado (ou previsto) dos dados quando eles seguem uma distribuição normal), x_{sw} (São os dados da amostra ordenados, utilizados no cálculo dos coeficientes necessários para realizar o teste)(Pino, 2014).

O Teste é definido matematicamente na Equação 2.11:

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad . \tag{2.11}$$

Onde:

- W é a estatística de teste de Shapiro-Wilk;
- n é o tamanho da amostra;
- $x_{(i)}$ é o i-ésimo menor valor ordenado;
- x_i é o i-ésimo valor observado;
- $\bar{x}$ é a média das observações;
- a_i são os coeficientes de correção, que dependem das médias e covariâncias das ordens dos valores.

2.7.3 *Bootstrapping*

O *bootstrapping* é uma técnica de reamostragem com reposição (processo de criar amostras a partir de um conjunto de dados original), que visa reduzir desvios e fornecer uma estimativa mais confiável da variância. A reamostragem, neste contexto, não acrescenta informações novas à amostra original, mas permite explorar as propriedades estatísticas dos dados disponíveis. Essa técnica não substitui dados com o objetivo de aumentar a precisão, mas serve para compreender melhor a distribuição de uma estimativa. A ideia central do *bootstrapping* é estimar a distribuição de uma estatística de interesse a partir de uma única amostra, gerando múltiplas amostras adicionais (com reposição) e, assim, avaliar a incerteza e construir intervalos de confiança para essa estatística. Esta técnica, é particularmente útil em situações, em que a formulação de modelos teóricos complexos para a variância é inviável.

O funcionamento do *bootstrapping* é da seguinte forma: seja $(y_1, y_2, ..., y_n)$ a amostra dada, retira-se dessa amostra uma amostra de tamanho 'n' com reposição. Essa amostra $B_j = (y_1^*, y_1^*, ..., y_n^*)$ é o *bootstrapping*, cada y_i^* é uma escolha aleatória de $(y_1, y_2, ..., y_n)$ para $j = 1, 2, .., m$. Então, para cada amostra, é calculado uma estatística de interesse β_j, que pode ser uma média, mediana, desvio padrão entre outras, dependendo do que esta tentando estimar.

A distribuição das estatísticas β_j, obtidas das amostras *bootstrap* forma a distribuição *bootstrap* do estimador θ, que é o parâmetro que estamos interessados em

estimar. Assim, θ pode representar a média da população, a mediana, ou qualquer outra medida que se deseje avaliar(da Silva Filho, 2010).

2.7.4 Análise de Variância - ANOVA

A análise de variância é uma ferramenta de comparação que investiga a existência de diferenças significativas entre os grupos estudados, inferindo um nível de significância. A ANOVA permite a comparação entre três ou mais grupos, sua aplicação é testar se existe diferença entre as médias das amostras. Seu conceito é baseado na soma dos quadrados como o início dos passos para obter a variação total, entre e dentro dos grupos. Em situações de comparações, pode-se verificar a soma total de quadrados, a soma entre grupos e a soma dentro dos grupos, que estão dispostos da seguinte forma nas equações 2.12, 2.13 e 2.14; respectivamente (Paese et al., 2001).

$$SQ_{total} = \sum X_{total}^2 - N_{total}\overline{X}_{total}^2 \quad . \tag{2.12}$$

$$SQ_{dentro} = \sum X_{total}^2 - \sum N_{grupo}\overline{X}_{total}^2 \quad . \tag{2.13}$$

$$SQ_{entre} = \sum N_{grupo}X_{grupo}^2 - N_{total}\overline{X}_{total}^2 \quad . \tag{2.14}$$

Existe uma medida de variação, na qual dividimos a soma dos quadrados por um número chamado “graus de liberdade” (representam o número de valores independentes usados para calcular as variâncias entre os grupos e dentro dos grupos) como mostra as equações 2.15 e 2.16.

$$MQ_{entre} = \frac{SQ_{entre}}{gl_{entre}} \quad . \tag{2.15}$$

$$MQ_{dentro} = \frac{SQ_{dentro}}{gl_{dentro}} \quad . \tag{2.16}$$

A análise dessas medidas de variação produz uma razão F, chamada de coefi-

ciente de *Fisher*, com o objetivo de determinar se há diferenças estatisticamente significativas nas médias dos grupos.

O numerador deste parâmetro indica a variação entre os grupos que são comparados e possui um denominador que armazena uma estimativa de variação dentro desses grupos. O coeficiente de *Fisher* é definido segundo a Equação 2.17.

$$F = \frac{MQ_{entre}}{MQ_{dentro}} \quad . \tag{2.17}$$

Dessa forma, a ANOVA indica o tamanho da média quadrática entre grupos relativos ao valor da média quadrática dentro dos grupos, isso indica que quanto maior a razão F, maior será a probabilidade de rejeitar a hipótese nula, sugerindo que pelo menos um dos grupos tem uma média, significativamente, diferente dos outros.

2.8 MÉTODOS DE AVALIAÇÃO

Este subtópico refere-se à definição de técnicas utilizadas para medir o desempenho dos resultados obtidos com este estudo.

2.8.1 Razão Sinal - Ruído de Pico (PSNR)

Método utilizado para medir a qualidade de uma imagem comprimida em relação a sua versão original ou de referência, muito utilizada na área de processamento digital de imagens e compressão de dados. Mas, neste estudo o método (PSNR) será utilizado para avaliar a qualidade da imagem binarizada ao ser comparada com a imagem original, A fórmula do PSNR é dada pela Equação 2.18.

$$PSNR = 10 * log_{10}(\frac{Maximo\ valor\ do\ pixel^2}{MSE}) \quad . \tag{2.18}$$

Em que,

- 'Máximo valor do *pixel*' é o valor máximo, que um *pixel* pode ter na imagem. Geralmente, 255 valores para imagens codificadas com 8 bits por *pixel* por

canal, ou seja, $2^8 = 256$ valores diferentes (0 à 255).

- MSE (Erro Médio Quadrático ou *Mean Squared Error*) é a média dos quadrados das diferenças *pixel* a *pixel* entre a imagem original e a imagem comprimida.

O PSNR é expresso em decibéis(dB) e fornece uma medida relativa de quanto a qualidade da imagem é preservada após o processo de filtragem. Quanto maior o valor do PSNR, melhor será a qualidade da binarização.

Erro Médio Quadrático - MSE

O Erro Médio Quadrático (MSE- *Mean Squared Error*) é um método utilizado para avaliar a diferença entre duas imagens. O MSE calcula a média dos quadrados das diferenças entre os valores dos *pixels* nas duas imagens, ou seja, ele se refere ao erro entre duas imagens F_A e F_B de dimensões (M x N), onde quanto menor o erro, mais semelhantes as imagens são. Assim o método permite uma interpretação direta de similaridade (Oliveira et al., 2019).O MSE entre dois sinais é dado pela Equação 2.19.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad . \tag{2.19}$$

- MSE é o Erro Médio Quadrático;
- n é número total de amostras nos dados;
- y_i é o valor real (observado) para a i-ésima amostra;
- $\hat{y}_i$ é o valor previsto pelo modelo para a i-ésima amostra.

Mapeamento de *pixels*

O mapeamento de *pixels* é o processo de associar cada *pixel* em uma imagem de entrada a um ou mais *pixels* correspondentes em uma imagem de saída. Esse procedimento é empregado em áreas, tais como, visão computacional, correspondência de

características (processo de identificar e alinhar pontos ou regiões semelhantes entre duas ou mais imagens), reconhecimento de padrões e registro de imagens, pois estabelece correspondências significativas. Neste trabalho, o mapeamento é utilizado para identificar os valores de cada *pixel* nas imagens binarizadas, a fim de comparar os resultados de modelos clássicos de binarização com o modelo proposto, em busca de semelhanças. Isso é feito calculando a porcentagem de semelhança, que corresponde à razão de pontos semelhantes sobre o total de pontos analisados.

2.9 *SOFTWARE* E AMBIENTES UTILIZADOS

Este subtópico serve para descrever as ferramentas e bibliotecas utilizadas para auxiliar o tema apresentado, como o MaZda®, Excel® e *Python*®.

2.9.1 *Software* MaZda®

MaZda® é utilizado para o cálculo de parâmetros de textura em imagens digitalizadas. Foi originalmente desenvolvido em 1996 no *Institute of Electronics da Technical University of Lodz (TUL)*, na Polônia, por *Michal Strzelecki* e *Piotr Szczypinski*, para análise de textura de mamografias. Seu nome é um acrônimo de *'Macierz Zdarzen'* que em polonês significa "matriz de eventos ou matriz de co-ocorrência". Este *software* permite o cálculo de uma variedade de parâmetros estatísticos que são derivados do histograma da imagem, gradiente absoluto (medida que indica a taxa de mudança máxima de intensidade de um pixel em uma imagem), matriz de co-ocorrência (tabela que mostra como diferentes valores de intensidade de pixel ocorrem juntos em uma imagem), matriz de comprimento de execução (representação que conta o número de pixels de cada valor de intensidade em uma imagem) e modelo autorregressivo (modelo estatístico usado para prever valores futuros com base em valores passados de uma série temporal). São cerca de 300 atributos gerados a partir da textura da imagem.

2.9.2 *Software* Excel®

Software de planilha amplamente utilizado para criação e análise de banco de dados, cálculos e criação de *dashboards* (painéis de controle interativos que permitem visualizar e analisar dados de maneira clara e concisa). Neste trabalho, foi utilizado como uma ferramenta de banco de dados para o *Python*®.

2.9.3 *Software Python*®

Python® é um linguagem de programação de alto nível(linguagem que se aproxima da linguagem humana, permitindo que desenvolvedores escrevam código de maneira mais intuitiva), interpretada, de propósito geral e bastante versátil. Com uma sintaxe clara, possui várias aplicações, tais como desenvolvimento web, análise de dados e aprendizado de máquina. O ambiente utilizado no estudo foi o *Jupyter Notebook* (interface interativa que permite escrever e executar código em um ambiente baseado na web), e as bibliotecas de *Python*® utilizadas foram:

- **Pandas**: para importar os dados do Excel®;
- ***Numpy***: para operações matemáticas;
- ***Seaborn***: para visualização de dados;
- ***matplotlib.pyplot***: para criação e visualização de gráficos e dados em geral;
- ***from scipy.stats import anderson***: pacote para importação de funções estatísticas, especificamente o teste de Anderson Darling.

Capítulo 3

METODOLOGIA

Este capítulo tem como objetivo orientar o leitor sobre as técnicas adotadas para o planejamento, condução e análise no desenvolvimento do trabalho, indicando a combinação de métodos quantitativos (são abordagens que utilizam dados numéricos e estatísticas para analisar e interpretar fenômenos) e qualitativos (são abordagens que exploram aspectos subjetivos e contextuais de fenômenos) para estudar a binarização de imagens. Essas técnicas abrangem desde a aquisição da base de dados até a comparação de métodos de binarização, dividindo o processo da seguinte forma:

- **Aquisição de Imagens:** Consiste em obter um banco de imagens, cartas históricas cedidas pela Fundação Joaquim Nabuco, para a conversão de imagem em dados.
- **Procedimentos para a extração de características:** Extrai informações de textura através da seleção das ROI's na imagem para obter o cálculo de atributos da imagem pelo *software* MaZda®.
- **Processamento de dados:** Organiza e trata dados a partir dos atributos selecionados, para realizar cálculos estatísticos. A técnica utilizada foi o teste de normalidade. Neste tópico também foram utilizados os *softwares*: Excel® e *Python*®, de forma a obter conclusões, por meio de ferramentas estatísticas, tais como, o teste de normalidade.

- **Análise e ajustes:** Interpreta os resultados obtidos no processamento de dados e propõe ajustes, como o *bootstrapping*.

- **Processamento de dados após às análises e ajustes:** Refaz o teste de normalidade após ajuste e aplica o teste de hipóteses (ANOVA), com o objetivo de comparar as médias dos grupos independentes. Este tópico também é responsável por encontrar o *threshold* estatístico.

- **Binarização:** Inseri o *threshold* estatístico ao modelo de binarização global e aplica as imagens históricas, que resulta na binarização de imagens

- **Avaliação:** Métodos clássicos de binarização foram executados, para avaliar o resultado do trabalho por processos de comparação, por meio de métricas como PSNR e o mapeamento de *pixels*.

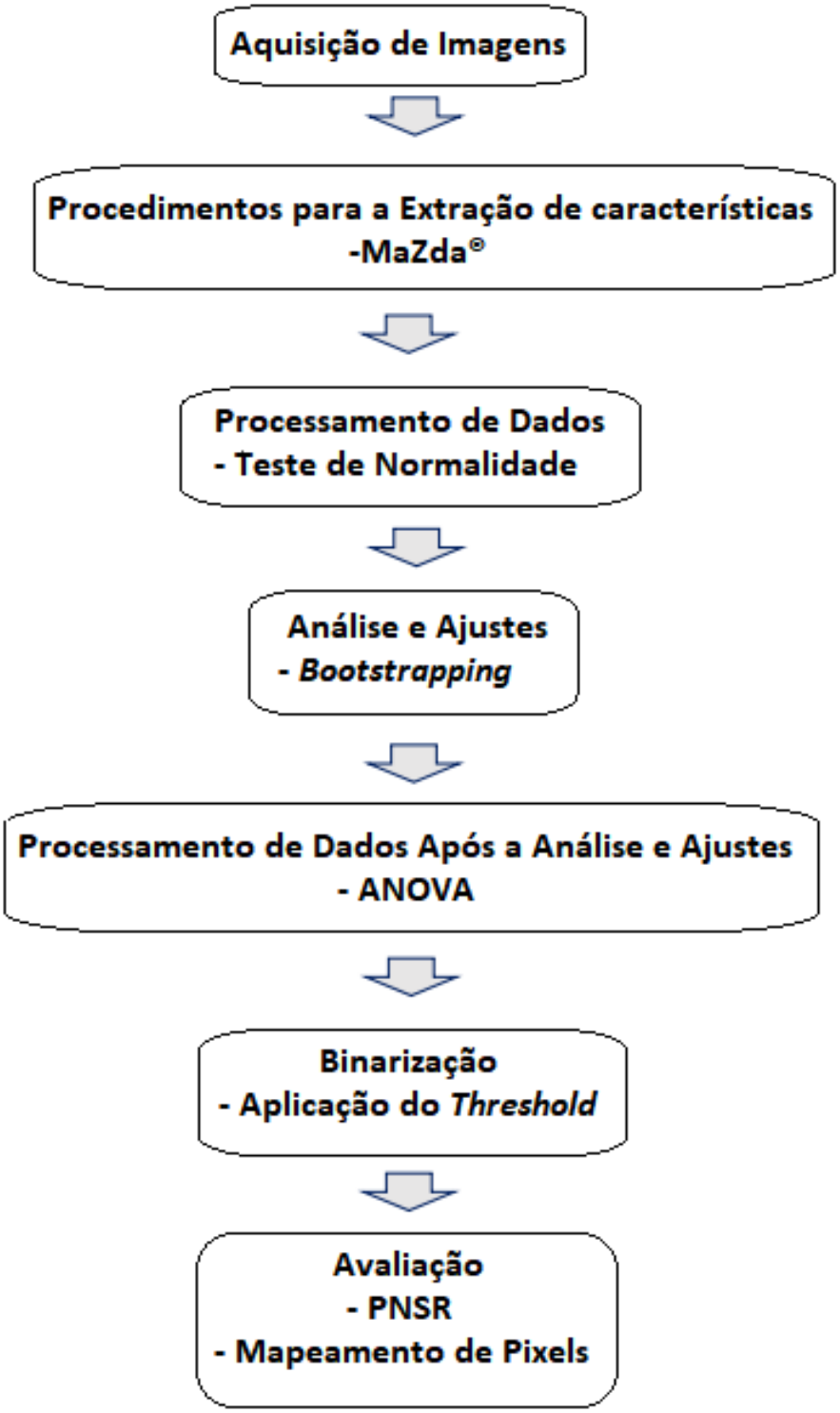
3.1 DIAGRAMA DE BLOCOS DO PROCESSO DE BINARIZAÇÃO

Para melhor caracterizar o processo de binarização estatística proposto, neste trabalho, foi criado um diagrama de blocos, que apresenta a ordem das ferramentas, teste estatísticos e os métodos avaliativos utilizados. Este diagrama pode ser visualizado na Figura 3.1.

3.2 AQUISIÇÃO DE IMAGENS

O primeiro passo da metodologia consiste em obter uma quantidade suficiente de imagens para a extração de dados (40 imagens, pois a partir dessas imagens é possível obter uma quantidade maior de dados). Essas imagens são cartas históricas digitalizadas, obtidas da Fundação Joaquim Nabuco. As cartas históricas foram escolhidas por apresentarem deterioração do papel, descoloração, manchas e ruídos identificados na frente e no verso. Isso tudo torna a leitura do texto uma tarefa difícil. A escolha dessas cartas deve-se ao desafio da melhora da legibilidade do texto.

Figura 3.1: Diagrama de blocos do processo de binarização

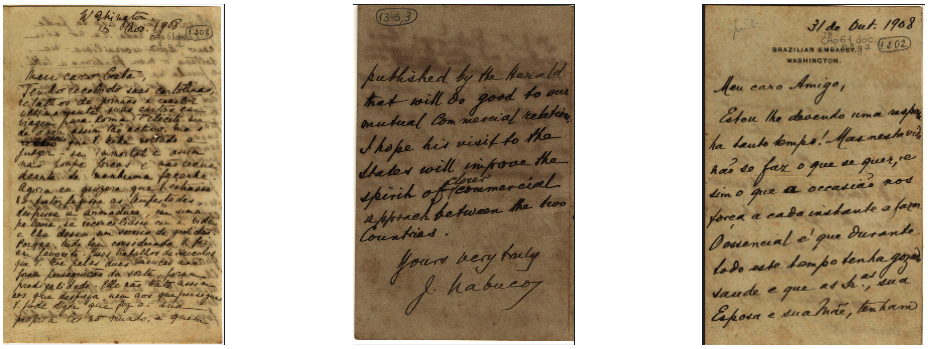


(Fonte: Autor da Figura)

No segundo passo, foi realizada uma seleção de imagens, com o objetivo de obter alta representação da escrita, forte envelhecimento do papel e baixa interferência de ruídos, tanto na frente quanto no verso. Desta forma, selecionou-se aleatoriamente, 40 imagens, como apresentadas nas Figuras 3.2(a), escolhida pela alta interferência de ruído na frente e verso da carta, 3.2(b), devido ao forte ruído por causa do envelhecimento do papel e 3.2(c), pela interferência média de ruídos, que passaram por um processo de extração de características.

Figura 3.2: Amostras da base de dados

(a) MCW01 **(b)** NAB03 **(c)** MCA01



(Fonte: Fundação Joaquim Nabuco, 2020)

3.3 PROCEDIMENTOS PARA A EXTRAÇÃO DE CARACTERÍSTICAS

Este procedimento envolve o processamento de imagens, com o objetivo de fornecer ferramentas para facilitar a identificação e extração de informações. Consiste em aplicar máscaras nas imagens e selecionar ROI's (Regiões de Interesse), aplicando o *software* MaZda®, a fim de obter características das imagens.

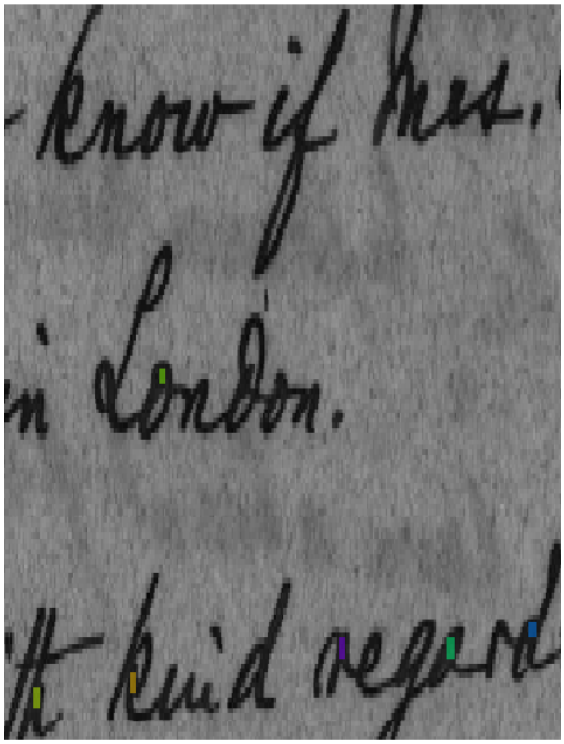
O MaZda® foi escolhido para a extração de características das imagens, devido à sua especialização e eficácia em pesquisas pontuais. Este *software* de tratamento de imagens foi aplicado a uma população de 40 imagens. Para cada imagem selecionou-se 16 ROI's, aleatórias, com o objetivo de obter características da escrita e do papel. Dessa forma, obteve-se 640 características (40 imagens * 16 ROI's= 640) para cada atributo, da amostra da letra e do papel. Isto resultou numa população de 1280

características (640 da amostra do papel + 640 da amostra da letra= 1280), para um único atributo.

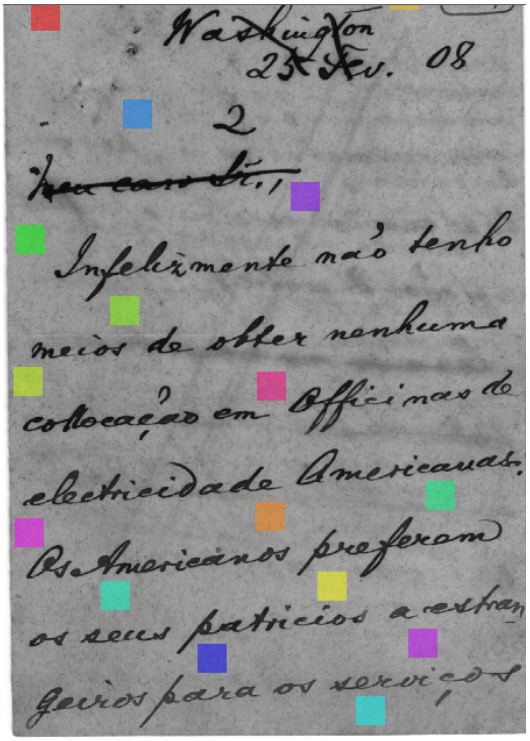
Para selecionar o tamanho das ROI's, foi definida uma janela, de forma experimental. O objetivo é de maximizar 16 regiões diferentes para fazer o processo de amostragem das amostras do papel e da escrita. Escolheu-se 50 x 50 *pixels*, para identificar as informações do papel, conforme mostrado na Figura 3.3(b). Para identificar as informações da escrita, foi determinada uma janela de 7 x 7 *pixels*, conforme ilustrado na Figura 3.3(a).

Figura 3.3: Amostragem feita na escrita e no papel

(a) 16 ROI's da escrita



(b) 16 ROI's do papel



(Fonte: Autor da Figura)

Dessa forma, obteve-se uma população total de 80 imagens (40 imagens, focadas em dois processos de amostragem, o que dobrou a respectiva base de imagens), dividida em 40 imagens referentes ao papel e as mesmas 40 imagens referentes à escrita. Essas imagens foram analisadas pelo *software* MaZda®, através da matriz de co-ocorrência, responsável por fornecer parâmetros, tais como média, variância, *skewness*(assimetria), curtose e outros. Os principais parâmetros estão definidos no

Anexo A.

Todo este processo de extração de características pelo *software* MaZda® é um processamento digital de imagens (PDI), cuja função é facilitar a identificação e extração da informações. Assim, o PDI deve ser encarado como uma fase preparatória para a atividade de interpretação das imagens (Crosta, 1999). Apesar da quantidade de parâmetros obtidos, o objetivo é selecionar aquele que melhor se adapta aos processos de binarização. Neste estudo, optou-se por utilizar a média dos tons de cinza como parâmetro.

3.4 PROCESSAMENTO DE DADOS

Nesta fase, os dados obtidos pelo *software* MaZda® são transferidos para o Excel®, que é utilizado como uma ferramenta auxiliar para importar as informações, para o ambiente de programação *Python*®. A escolha do *Python*® se deu devido à sua sintaxe clara e à disponibilidade de diversas bibliotecas específicas para análise de dados.

O trecho do algoritmo a seguir, Figura 3.4,descreve a importação dos dados do Excel® para *Python*®, bem como as bibliotecas que contribuíram para o desenvolvimento do estudo.

```
1 import pandas as pd
2 import numpy as np
3 import seaborn as sns
4 import scipy.stats as scs
5 import matplotlib.pyplot as plt
6 from scipy.stats import normaltest
7 P=pd.read_csv('Pok.csv',sep=';')
8 L=pd.read_csv('LoksNaN.csv',sep=';')
```

Figura 3.4: Quadro 1 - Algoritmo da Importação do dataset

Após a importação dos dados, o próximo passo foi aplicar o histograma e o gráfico de quantis (QQ-Plot), gráfico baseado no escalonamento dos dados pelo des-

vio padrão às amostras. O objetivo foi visualizar os dados e auxiliar na identificação de desvios da normalidade e na compreensão da conformidade dos dados com uma distribuição normal. Essas representações gráficas são essenciais para uma análise inicial e para fundamentar decisões em testes subsequentes.

Após às análises gráficas, foi realizado o teste de normalidade de Anderson-Darling, para avaliar se a hipótese nula de uma distribuição de dados segue a normalidade. Com base nisso, formulou-se às seguintes hipóteses:

- Hipótese nula (H0): a amostra segue a distribuição normal;
- Hipótese Alternativa (H1): a amostra não segue a distribuição normal.

O resultado dos testes de normalidade indicou que tanto os dados do papel quanto os da escrita não seguem uma distribuição normal. Isso ocorreu porque o valor da estatística de Anderson-Darling foi menor, que o valor crítico, o que leva à rejeição da hipótese nula.

O trecho do algoritmo abaixo, Figura 3.5,descreve a execução do teste de Anderson-Darling.

```
from scipy.stats import anderson
A1 = anderson(P["Mean"])
#A1
E=round(A1.statistic,3)
C=round(A1.critical_values[2],3)
#Se o valor da estat stica for menor que o valor cr tico, os
    dados seguem a normal
if A1.statistic<A1.critical_values[2]:
    print("Os dados do papel Seguem a normalidade, pois o valor do
    par metro estat stico", E ,"  Menor que o par metro cr tico
    ", C, ".")
else:
    print("Os dados do papel  N  O  Seguem a normalidade, pois o
    valor do par metro estat stico", E ,"  MAIOR que o par metro
     cr tico", C, ".")
```

Figura 3.5: Quadro 2 - Algoritmo do Teste de Anderson Darling

3.5 ANÁLISE E AJUSTES

Nesta etapa, foi realizado uma análise baseada no teste de normalidade, pois os dados obtidos no estudo proposto não seguiam uma distribuição normal. Então, realizou-se a técnica chamada *bootstrapping*, para gerar amostras comparáveis com a finalidade de se obter, uma significância estatística dos dados (a normalidade). Esta técnica demonstrou que quanto maior o número de amostras, mais preciso é o intervalo de confiança. Portanto, foram definidas 10.000 amostras, com o objetivo de garantir um intervalo de confiança mais estável. Após a aplicação dessa técnica, o teste de normalidade foi refeito resultando em dados normalmente distribuídos, tornando-os adequados para análises paramétricas.

O trecho do algoritmo da Figura 3.6, descreve a execução do *bootstrapping*.

```
def drawNewSample(data,func1,func2):
    sample=np.random.choice(data,size=len(data))
    return func1(sample),func2(sample)
arrayDataP=P['Mean']
n1=len(arrayDataP)
#n1
meanOrigSample1=np.mean(arrayDataP)
stdOrigSample1=np.std(arrayDataP)
B1=10000 # 10000 VALORES DE REAMOSTRAS EXTRAIDAS DA AMOSTRA
   ORIGINAL
arrayM1=np.empty(B1)#vetor com medias
arrayStd1=np.empty(B1)#vetor com desvio padrao
for i in range(B1):
    vm1,vdp1=drawNewSample(arrayDataP,np.mean,np.std)
    #CHAMA A FUNCAO QUE FAZ O SORTEIO ALEATORIO
    arrayM1[i]=vm1
    #VALOR COM "MAIOR FREQUENCIA" OU "MEDIANA" E O VALOR QUE EU
   UTILIZO COMO threshold EM OTSU
    arrayStd1[i]=vdp1
    #print(f'Amostra {i+1}: {arrayM1[i]:.2f} {arrayStd1[i]:.2f}')
RP = []
for i in range(B1):
    RP.append(arrayM1[i])
```

Figura 3.6: Quadro 3 - Algoritmo do *Bootstrapping*

3.6 PROCESSAMENTO DE DADOS APÓS ÀS ANÁLISES E AJUSTES

O teste paramétrico adotado neste trabalho foi a análise de variância (ANOVA), com o objetivo de comparar as médias dos grupos independentes e identificar diferenças significativas entre as médias populacionais. Esse teste foi realizado em duas bases de dados: a base do papel e a da escrita. Após o processo de *bootstrapping*, obteve-se uma população de 20.000 dados, 10.000 para o papel e 10.000 para a escrita. As amostras foram trabalhadas separadamente. Para cada grupo de 10.000 amostras, foi separadas 10 grupos de 1.000 amostras independentes. O objetivo é identificar diferenças significativas entre as médias dos grupos independentes, tanto da amostra do papel como da amostra da letra. A ANOVA é baseada nas seguintes hipóteses:

- Hipótese nula (H0): as médias dos grupos são iguais;
- Hipótese alternativa (H1): pelo menos uma média é diferente das outras

O trecho do algoritmo abaixo, descreve a execução da análise de variância.

```
1 from scipy.stats import f_oneway
2 #perform one-way ANOVA
3 f_oneway(LL1, LL2, LL3,LL4, LL5, LL6,LL7, LL8, LL9,LL10)
```

Figura 3.7: Quadro 4 - Algoritmo da ANOVA

3.7 BINARIZAÇÃO

Nesta etapa foi determinado o *threshold* e em seguida aplicado o algoritmo de binarização global. Este valor é obtido pela média global dos níveis de tonalidade de cinza, que é a média das médias locais de cada amostra, no caso da escrita e do papel. De posse do *threshold* obtido e posteriormente aplicado ao algoritmo, o programa obteve a binarização da imagem, que processou uma carta histórica

com muitos ruídos em uma imagem binarizada em preto e branco, onde é possível distinguir a escrita do papel.

O estudo também proporcionou uma análise de sensibilidade ao valor do *threshold* obtido, por meio da variação de 30%, 50%, 60%, 70% de incremento e decremento. O intuito foi analisar, a partir de qual ponto, ocorre, uma variação brusca na binarização. No método de binarização proposto (global), a imagem é convertida em tons de cinza e, posteriormente, em uma imagem binária, ao utilizar um valor chamado *threshold*.

Neste trabalho, a determinação do *threshold* foi feito de várias maneiras: pela análise do histograma (visualmente), pela utilização métodos estatísticos (proposta do trabalho) e no ajuste do algoritmo da binarização global (forma empírica) de forma dinâmica, como no caso das variações percentuais do *threshold*. A forma que será evidenciada será a utilização de métodos estatísticos. No Anexo C são apresentados algoritmos que executam as binarizações adaptativas e globais.

3.8 AVALIAÇÃO

Métodos clássicos de binarização, como Otsu, adaptativa pela média, adaptativa pela Gaussiana, *Niblack* e Sauvola, foram utilizados para comparar tanto visualmente, como analiticamente, com o método proposto, baseado na estatística.

A binarização adaptativa foi realizada com o objetivo de analisar uma binarização mais eficaz, dividindo a imagem em blocos nos quais um *threshold* é calculado para cada bloco individualmente. Isto permite, considerar as variações de iluminação na imagem. A binarização de Otsu foi utilizada, pois é um método global e uma forma de comparar resultados de *thresholds* obtidos de bibliotecas do *software* e estatisticamente.

Os processos de comparação de resultados levaram em consideração duas métricas o PSNR, relação sinal-ruído de pico, e o mapeamento de *pixels* das imagens binarizadas. O PSNR é uma métrica utilizada para avaliar a qualidade de uma imagem com base na relação do sinal original e o ruído. Dessa forma, foram comparadas as

imagens que passaram por cada tipo de binarização, com o objetivo de selecionar os filtros com maior PSNR, pois quanto maior o PSNR. Dado que, quanto maior o PSNR melhor será o modelo empregado. O mapeamento de *pixels* foi realizado com a finalidade de identificar as binarizações que mais se aproximam da proposta do trabalho. Para isso, comparou-se o mapeamento feito por modelos clássicos com a binarização proposta por meios estatísticos. Assim comparou-se as imagens binarizadas *pixel* a *pixel*. A quantidade de pontos semelhantes foi obtida e dividida pela quantidade de *pixels* analisados. Dessa forma, foi possível, obter uma porcentagem de pontos semelhantes ao modelo estatístico proposto.

Capítulo 4

RESULTADOS

Neste capítulo, são abordados os resultados analíticos e gráficos obtidos a partir da solução proposta, bem como a comparação com outros métodos de binarização mencionados. O objetivo é avaliar a eficácia dos resultados do modelo proposto em relação a essas alternativas. Na seção 4.1, são apresentados os resultados gráficos e estatísticos da obtenção do *dataset*(conjunto organizado de dados, estruturado em tabelas ou arquivos, usados para análise) e do *threshold*, os quais foram obtidos por meio dos procedimentos de extração de características, processamento de dados, análise e ajustes, assim como o processamento posterior à análise e aos ajustes. Na seção 4.2, são apresentados os resultados do modelo proposto e sua comparação com binarizações clássicas. Na seção 4.3, são apresentados os resultados analíticos, tais como a Relação Sinal-Ruído de Pico (PSNR) e o Mapeamento de Pontos em Imagens, os quais foram utilizados para avaliar os melhores modelos.

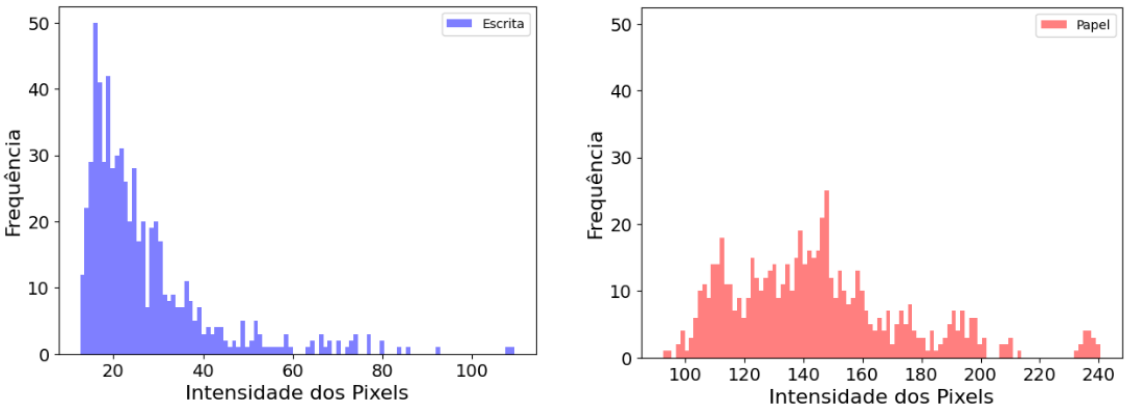
4.1 VISUALIZAÇÃO DOS RESULTADOS GRÁFICOS ESTATÍSTICOS

Nesta etapa do processo de binarização, as imagens passaram por um procedimento de extração de características realizados, pelo MaZda®, que resultou em uma grande quantidade de atributos estatísticos de 1º e 2º ordem. O atributo escolhido para desenvolver a proposta do trabalho, foi a média dos tons de cinza do papel e

da escrita, estes dados são apresentados segundo os histogramas da Figura 4.1, com a representação da escrita na Figura 4.1(a) e a representação do papel na Figura 4.1(b)

Figura 4.1: Histograma da distribuição dos dados da Escrita e Papel

(a) Escrita **(b)** Papel



(Fonte: Autor da Figura)

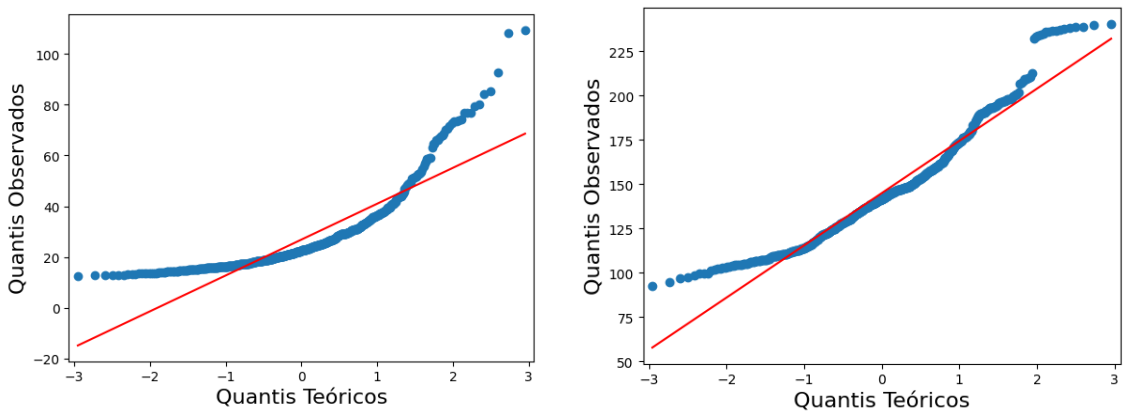
Os histogramas das Figuras 4.1(a) e 4.1(b) apresentam uma distribuição de dados com um formato diferente de uma curva Gaussiana, do formato sino. Então para verificar a disposição dos dados e comparar com uma distribuição normal, aplicou-se um QQ-plot (gráfico quantil-quantil). O gráfico QQ-plot foi usado para comparar a distribuição dos dados da textura dos tons de cinza do papel e da escrita (os dados utilizados foi os 1280 dados, 640 do papel e 640 da escrita ambos obtidos após o processo de extração de características pelo MaZda®), com uma distribuição teórica, como a distribuição normal (Gaussiana). O QQ-plot possui uma linha de referência baseada no escalonamento dos dados pelo desvio padrão, tendo a sua média adicionada a linha de referência. Dessa forma, quando os dados estão dispostos no QQ-plot e seguem essa linha de referência, os dados seguem a distribuição normal, caso não siga essa linha, haverá uma divergência com o modelo normal. Os QQ-plots gerados são apresentados segundo as Figuras 4.2(a) e 4.2(b), representando respectivamente a escrita e o papel.

Observou-se na Figura 4.2(a), que os dados da escrita divergem da linha de referência, concluindo um distanciamento da normalidade dos dados nesta amostra.

Figura 4.2: QQ-Plot da distribuição dos dados da Escrita e do Papel.

(a) Escrita

(b) Papel



(Fonte: Autor da Figura)

Na Figura 4.2(b), observou-se, que os dados do papel se aproximavam da linha de referência, linha vermelha, no intervalo dos quantis teóricos de -1 à 2, concluindo uma aproximação da normalidade dos dados do papel neste intervalo.

4.1.1 Análise e Ajustes

Apesar dos resultados gráficos obtidos pela distribuição dos dados, foi necessário realizar o teste de normalidade de forma analítica. O teste escolhido foi o de Anderson Darling, devido ao afastamento dos pontos no QQ-plot das amostras em relação à reta de referência, especialmente nas extremidades, pois sugere desvios nas caudas da distribuição, como mostra as Figuras 4.2(a) e 4.2(b). O teste de normalidade mostrou que tanto a distribuição dos dados da amostra do papel quanto da escrita não seguem uma distribuição normal, pois os parâmetros estatísticos são maiores que os parâmetros críticos, como são apresentados na Tabela 4.1.

Tabela 4.1: Resultado do Teste de Normalidade - Anderson Darling.

Amostras	Parâmetro Estatístico	Parâmetro Crítico	Resultado
Papel	8,272	0,782	Não normal
Escrita	41,099	0,782	Não normal

Como as distribuições dos dados do papel e da escrita não seguem a normalidade,

aplicou-se um método não paramétrico (técnica estatística que não faz suposições rígidas sobre a distribuição de dados, como a normalidade), chamado *bootstrapping*. O objetivo do *bootstrapping* foi gerar amostras de dados que mantivessem o padrão estatístico original, com a finalidade de alcançar a significância estatística dos dados, a normalidade. A geração de amostras de dados, aplicado pelo método não paramétrico, *bootstrapping*, como visto anteriormente, segue o seguinte processo: das 640 amostras do papel que passaram pela técnica *bootstrapping*, obteve-se uma base de dados de 10.000 dados (amostras do procedimento do papel), o mesmo foi realizado para a amostra da escrita. O elevado número da base de dados das amostras é devido a manter uma confiança na significância dos dados. O método *bootstrapping* aproximou a distribuição da média de um levantamento estatístico (média ou mediana), pois utilizou reamostragem com reposição de uma única amostra para gerar várias novas amostras. Isso permite calcular a mesma estatística, como média ou variância, repetidamente. A distribuição resultante dessas estatísticas reamostradas reflete a variação nos dados da amostra original, pois fornece uma estimativa da incerteza associada à estatística. Esses levantamentos (média e mediana) geralmente envolvem parâmetros, que são características numéricas, que descrevem a população, como média, variância ou proporção. Assim, o *bootstrapping* permite estimar a distribuição da média desses parâmetros sem fazer suposições sobre a distribuição original dos dados. Em resumo, o bootstrapping aproxima a distribuição de parâmetros de um levantamento estatístico, ao reamostrar dados, da amostra original, possibilitando a avaliação da variabilidade e incerteza nas estimativas.

Após a técnica do *bootstrapping*, realizou-se um, novo teste de Anderson-Darling para verificar a normalidade dos dados, que passaram a ser, agora, de 10.000 amostras, para cada elemento de estudo, papel e escrita. O novo teste confirmou, uma distribuição normal em cada amostra, conforme mostrado na Tabela 4.2. Percebe-se pela respectiva tabela, que o parâmetro crítico é maior que o estatístico, resultando em distribuições que seguem a normalidade.

Tabela 4.2: Resultado do Teste de Normalidade após o *bootstrapping*.

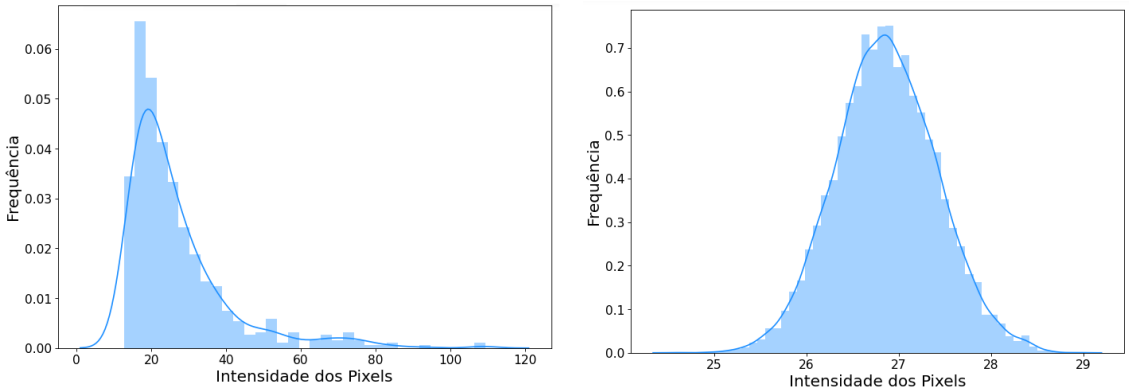
Amostras	Parâmetro Estatístico	Parâmetro Crítico	Resultado
Papel	0,344	0,787	normal
Escrita	0,511	0,787	normal

Nas Figuras 4.3(a), 4.3(b), 4.4(a) e 4.4(b) foi possível visualizar os histogramas das amostras do papel e da escrita antes e depois do *bootstrapping*, respectivamente.

Figura 4.3: Histograma dos dados referente a escrita, antes e depois do *Bootstrapping*

(a) Escrita - Antes do *Bootstrapping*

(b) Escrita - Depois do *Bootstrapping*

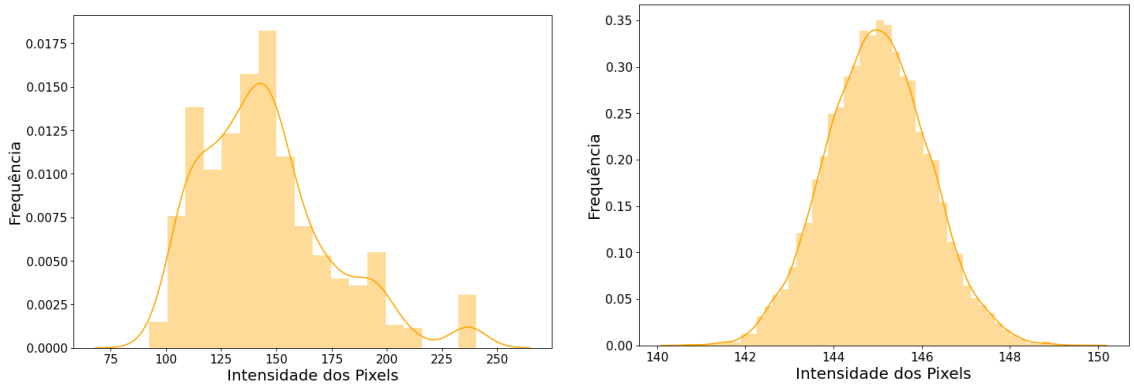


(Fonte: Autor da Figura)

Figura 4.4: Histograma dos dados referente ao Papel, antes e depois do *Bootstrapping*

(a) Papel - Antes do *Bootstrapping*

(b) Papel - Depois do *Bootstrapping*



(Fonte: Autor da Figura)

Na Figura 4.3(a) é possível verificar um histograma da escrita antes do método *bootstrapping*, com o comportamento de uma assimetria positiva (distribuição com cauda mais longa à direita), com uma curtose positiva (distribuição com caudas mais

longas e mais pontiaguda do que uma distribuição normal), achatando um pouco a distribuição da amostra.

No histograma da Figura 4.4(a) é possível observar uma assimetria positiva na amostra do papel antes da aplicação do *bootstrapping*, ou seja, uma cauda alongada para a direita, com uma curtose positiva, mas com os dados mais distribuídos. Mas, após a aplicação do método, ambos os dados das amostras, papel e escrita, se comportaram como uma curva de Gauss, como mostra as Figuras 4.3(b) e 4.4(b). Isso significa que os dados se distribuem simetricamente em torna da média e a frequência de valores diminui gradualmente conforme se afastam da média, formando uma curva em formato de sino. Assim, o resultado do *bootstrapping* resultou na significância dos dados, tornando a média uma medida aplicável, que será utilizada para o cálculo do *threshold* estatístico, que será visto mais adiante.

Análise de Variância - ANOVA

Após a constatação da normalidade dos dados, realizou-se a análise de variância, com o objetivo de comparar as médias de grupos independentes, determinando se há diferença estatística significativa entre as médias populacionais. Dessa forma, foi separada a base de dados definida pela re-amostragem em 10 grupos de 1.000 amostras em cada grupo, para cada amostra (papel e letra, trabalhadas separadamente). Para a análise de variância, definiu-se hipóteses, afim de verificar a existência de diferença entre as amostras. As hipóteses são definidas da seguinte forma:

- H_0: Não existe diferença entre o atributo escolhido (média dos tons de cinza) das amostras referentes as imagens destinadas para a análise do papel e da escrita.
- H_1: Há pelo menos um atributo (média dos tons de cinza) diferente das amostras referentes as imagens destinadas para a análise do papel e da escrita.

Os resultados da análise seguem a Tabela 4.3:

Tabela 4.3: Resultado da Análise de Variância - ANOVA.

Amostras	Estatística do teste F	p-value
Papel	0,6561	0,7494
Escrita	0,8594	0,5611

Como o valor p não é inferior a 0,05 (nível de significância utilizado para determinar se os resultados são estatisticamente significativos), não é rejeitada a hipótese nula. Significando que não há evidências estatísticas suficientes para afirmar que existe uma diferença significativa nas 10 amostras, tanto do papel como da escrita. Em termos práticos, o resultado da ANOVA indica que as variações observadas nas médias dos tons de cinza podem ser atribuídas ao acaso e não a uma diferença real entre os grupos. Portanto, a hipótese de que não existe diferença significativa entre as médias dos tons de cinza das duas amostras, permanece válida, e não se tem suporte estatístico para afirmar que existem diferenças significativas entre elas. Indicando que, com os dados disponíveis, a diferença em cada amostra não é grande o suficiente para ser detectada significativa.

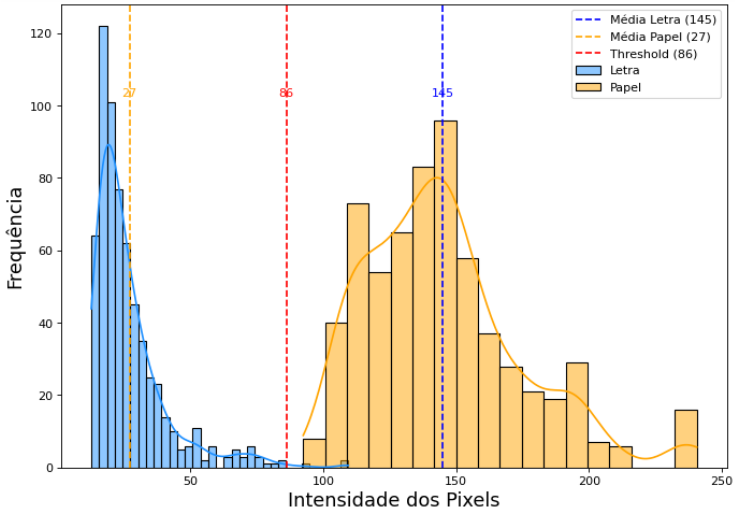
4.1.2 Cálculo do *Threshold* Estatístico

Ao finalizar os testes estatísticos, realizou-se as médias amostrais, que representam as médias do papel $\mu_P = 144,99 \approx 145$ e da escrita $\mu_E = 26,85 \approx 27$. Essas duas médias são chamadas médias locais, que obtendo a média delas é obtido a média global, $\mu_G = \frac{\mu_P + \mu_E}{2} = T = 86$, também chamado de *threshold*, esses dados podem ser visualizado no histograma da Figura 4.5. A obtenção deste valor, é de fundamental importância para o procedimento de binarização de imagens pelo modelo proposto.

4.2 RESULTADOS GRÁFICOS

Nesta etapa, são apresentados os resultados obtidos com os modelos de binarização, tanto do modelo proposto quanto do modelo clássico. Inicialmente, é apre-

Figura 4.5: Histograma dos dados referente ao Papel e a Escrita, antes do *Bootstrapping*, com o *threshold*

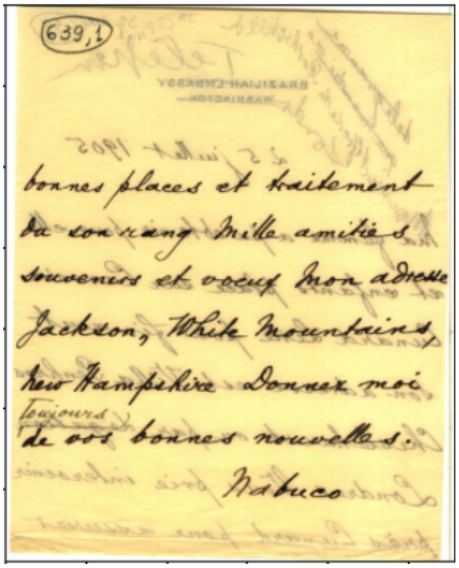


(Fonte: Autor da Figura)

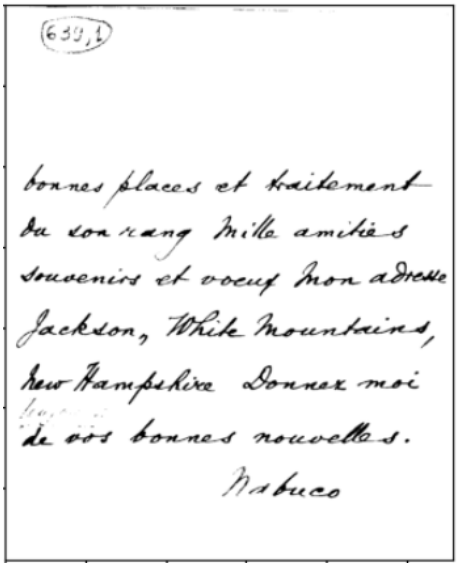
sentado o modelo baseado no estudo estatístico, no qual o *threshold* foi aplicado ao algoritmo de binarização global proposto, resultando uma imagem binarizada, como mostra a Figura 4.6(b). Na respectiva Figura 4.6(b), pode-se notar que o processo de binarização foi capaz de eliminar os ruídos da frente e do verso do documento da Figura 4.6(a), bem como removeu os efeitos do envelhecimento do papel e e caracterizou melhor a escrita.

Figura 4.6: Binarização baseado no modelo proposto

(a) Original NAB01



(b) Proposto NAB01, T- *Threshold* = 86



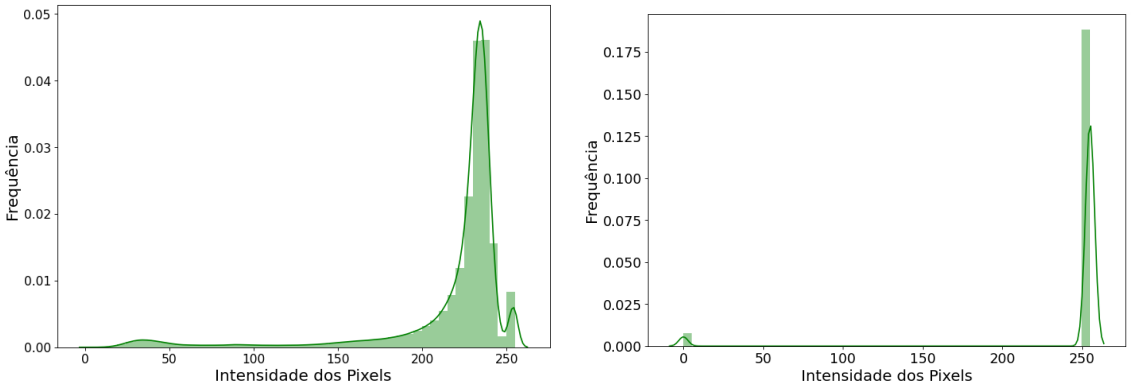
(Fonte: Autor da Figura)

Uma forma de evidenciar o resultado obtido, é por meio do histograma, que mostrou a frequência de cada tom separado entre dois grupos, que têm como referência o zero e o 255. Na Figura 4.7(a) pode-se ver o histograma da imagem antes do processo de binarização e na Figura 4.7(b), depois do processo. Identificando 2 picos, referentes as quantidades de *pixels* , um em 0 e outro em 255 que corresponde a 1.

Figura 4.7: Histograma da Imagem NAB01, antes e depois do processo de binarização.

(a) Original NAB01

(b) Proposto NAB01, T- *Threshold* = 86



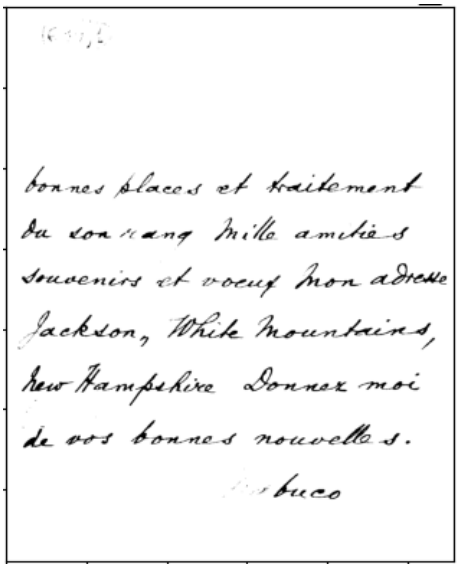
(Fonte: Autor da Figura)

Após o resultado validado pelo histograma foi realizado um decréscimo e um acréscimo gradual do *threshold*, uma análise de sensibilidade, $T \pm \sigma^2 = T_{\sigma^2}$ (T é o *threshold*, σ^2 é a variância e T_{σ^2} é o *threshold* com a variância) , com o intuito de obter uma faixa de legibilidade. Com o decréscimo gradual pode-se observar uma legibilidade com até 50% ($T - \sigma^2 = 86 - 43 = 43 = T_{50\%-}$) do *threshold*. Pode ser observado, também, que só a partir de um decréscimo de 30% ($T - \sigma^2 = 86 - 25,8 = 60.2 = T_{30\%-}$, uma diferença visual pode ser notada. Perde-se, totalmente, a legibilidade quando um decréscimo de 70% ($T - \sigma^2 = 86 - 60.2 = 25.8 = T_{70\%-}$) é aplicado, enquanto, o decréscimo de 60% é colocado para efeito de transição de resultados; As ilustrações dos decréscimos citados, podem ser vistas nas Figuras 4.8(a), 30%, 4.8(b), 50%, 4.8(c), 60% e 4.8(d), 70% de T.

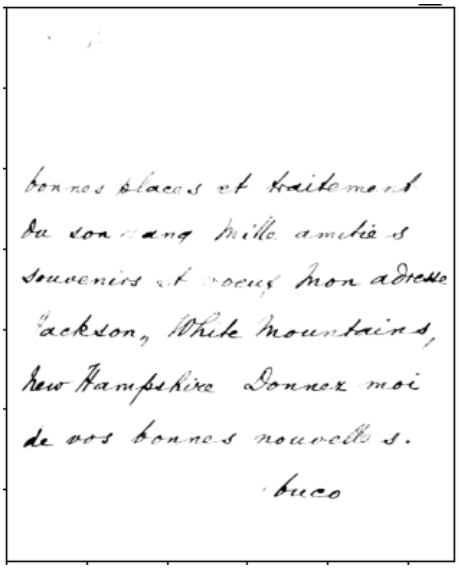
Ao realizar-se o acréscimo gradual do *threshold*, pode ser observado uma diferença visual no resultado a partir de 30% ($T+\sigma^2 = 86+25.8$), já em 50% ($T+\sigma^2 = 86+43$) a imagem binarizada começa a apresentar ruídos. Desta forma, pode-se confirmar

Figura 4.8: Resultado da análise de sensibilidade da binarização global aplicada ao modelo proposto, com variações negativas do *threshold*

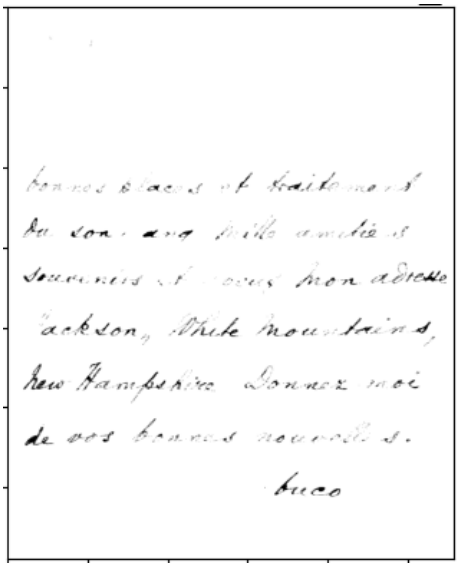
(a) $T_{0,7}$= 86 - 25.8 = 60.2



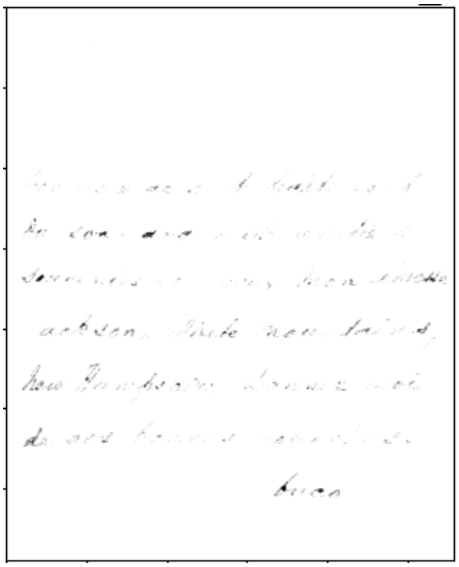
(b) $T_{0,5}$= 86 - 43 = 43



(c) $T_{0,4}$= 86 - 51.6 = 34.4



(d) $T_{0,3}$= 86 - 60.2 = 25.8

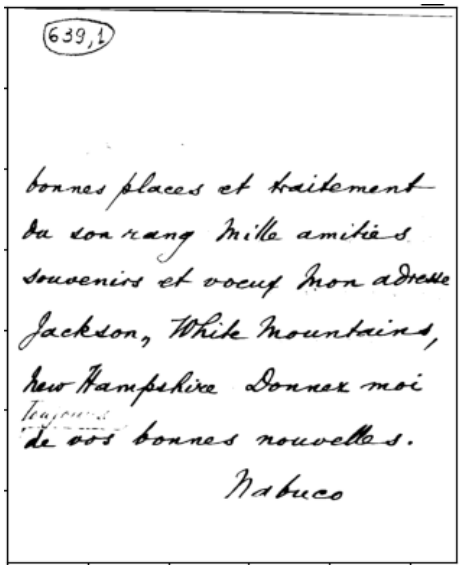


(Fonte: Autor da Figura)

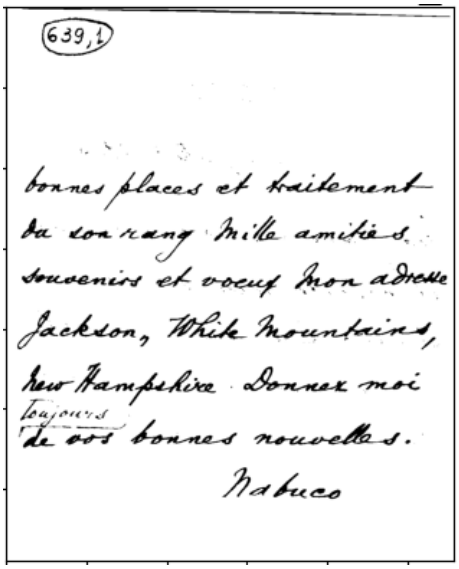
que o acréscimo do *threshold* endossa o ruído na binarização da imagem. As ilustrações com este acréscimo podem ser observadas nas Figuras 4.9(a), 30%, 4.9(b), 50%, 4.9(c), 60% e 4.9(d), 70%. Com essas variações de acréscimo e decréscimo aplicadas no valor obtido de *threshold*, foi obtido uma faixa de legibilidade do documento de 30% acima e abaixo do valor, que corresponde a uma faixa entre 66,15 e 111,80 para o valor do *threshold*.

Figura 4.9: Resultado da análise de sensibilidade da binarização global aplicada ao modelo proposto, com variações positivas do *threshold*

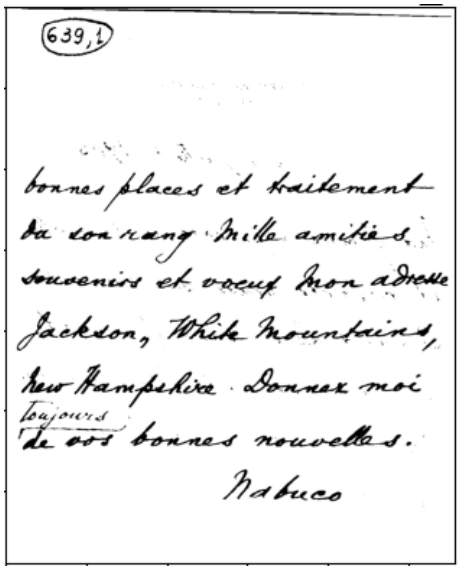
(a) $T_{1,3}$= 86 + 25.8 = 111.8



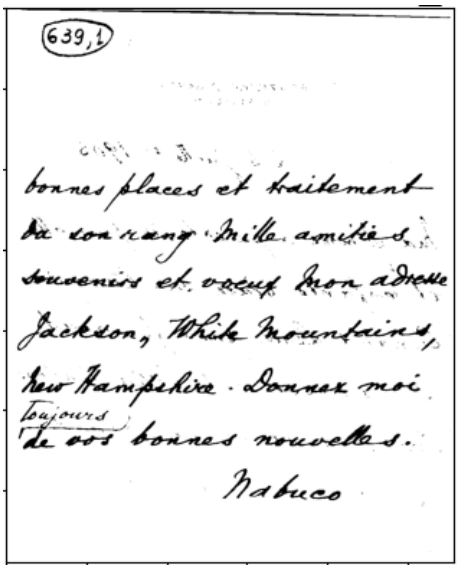
(b) $T_{1,5}$= 86 + 43 = 129



(c) $T_{1,6}$= 86 + 51.6 = 137.6



(d) $T_{1,7}$= 86 + 60.2 = 146.2



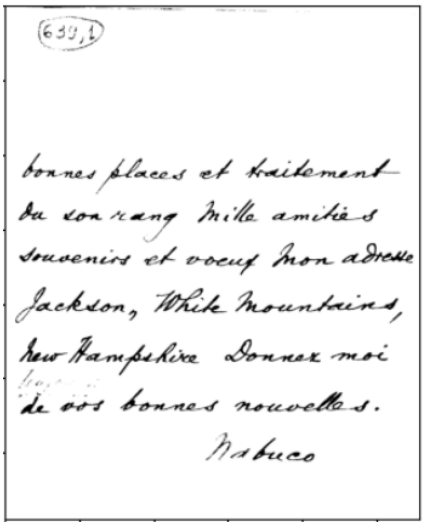
(Fonte: Autor da Figura)

Para efeitos de análise, realizou-se a binarização em modelos como Otsu, Sauvola,

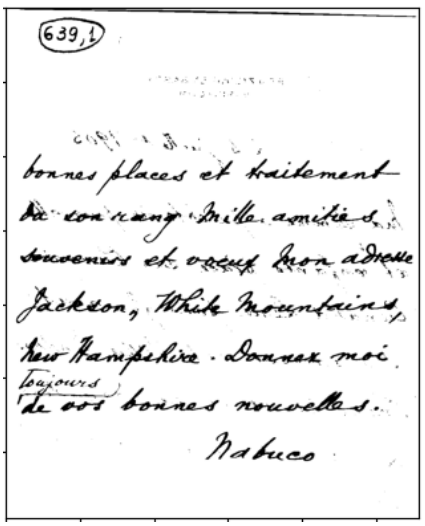
Niblack, Adaptativo pela Média, Adaptativo pela Gaussiana, com o objetivo de comparar os resultados dos modelos citados, com o resultado obtido pela binarização global, por meio do *threshold* estatístico. Entre os modelos automáticos, o Otsu é o que detém o melhor resultado, pois consegue filtrar a imagem, de forma a obter um resultado mais legível. Na Figura 4.10(a), pode-se observar, que o modelo proposto, estatisticamente, torna-se melhor, que o modelo automático de Otsu, Figura 4.10(b), pois o método proposto consegue filtrar com maior precisão, os ruídos.

Figura 4.10: Comparação das binarizações do modelo Proposto e o modelo automático de Otsu

(a) Proposto NAB01, T- *Threshold* = 86



(b) Otsu NAB01, T- *Threshold* = 151



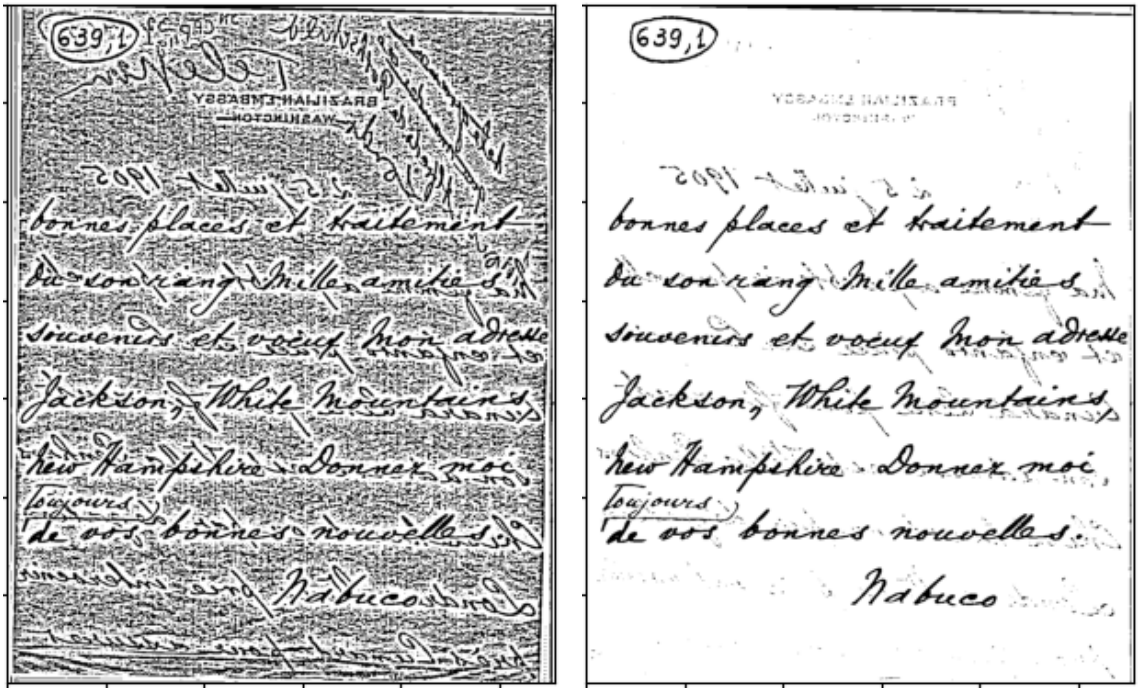
(Fonte: Autor da Figura)

Os modelos *Niblack* e Sauvola podem ser vistos respectivamente nas Figuras 4.11(a) e 4.11(b). O *Niblack* obteve uma grande quantidade de ruídos, já o Sauvola resultou como um modelo legível. Apesar desses resultados o Otsu ainda consegue ser melhor que o Sauvola.

As binarizações que tiveram o resultado mais perto do *threshold* estatístico foram os modelos adaptativos, que se baseiam em aspectos estatísticos de Média e da Gaussiana, demonstrados nas Figuras 4.12(a), 4.12(b).

Figura 4.11: Resultado das binarizações pelos modelos *Niblack* e Sauvola

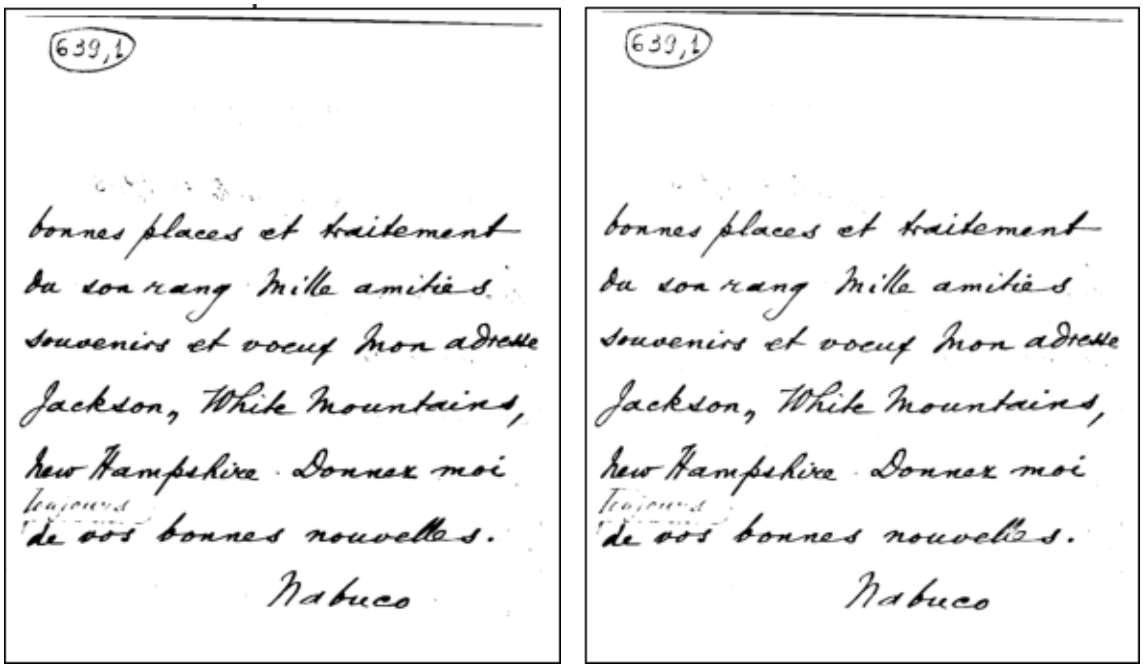
(a) *Niblack* NAB01 **(b)** Sauvola NAB01



(Fonte: Autor da Figura)

Figura 4.12: Resultado das binarizações Adaptativas pelos modelos da Média e da Gaussiana

(a) Adaptativo pela Média NAB01 **(b)** Adaptativo pela Gaussiana NAB01



(Fonte: Autor da Figura)

Após todos esses resultados, pode-se perceber visualmente que o melhor resultado converge para o modelo baseado pelo *threshold* tendo uma pequena faixa de variação, que foi observada na análise de sensibilidade pelo o acréscimo e decréscimo gradual do *threshold*.

4.2.1 Aplicação dos Métodos em Outras Amostras

Figura 4.13: Amostras Originais utilizadas para avaliação analítica

(a) NAB01 **(b)** PRES01 **(c)** MCA01

(d) EBX01 **(e)** ESC01 **(f)** NAB02

(Fonte: Fundação Joaquim Nabuco, 2020)

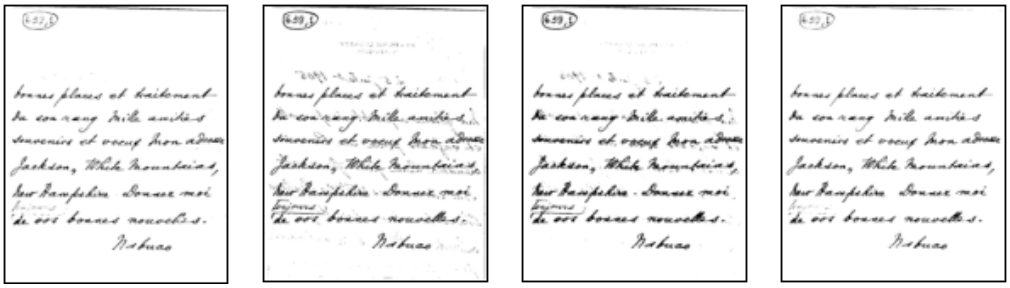
As amostras visualizadas da Figura 4.13, foram aplicadas em todos os métodos de binarização citados no estudo, com o objetivo de comparar os modelos e avaliar se o modelo proposto se aproxima ou se torna melhor que os modelos clássicos de binarização. Estas amostras foram escolhidas do banco de dados, com base em diferentes tipos de degradação do documento, como diferentes níveis de interferência de ruído nos caracteres do verso do papel e dos níveis de envelhecimento do papel.

As Figuras 4.13(a), 4.13(b), 4.13(c), 4.13(d) 4.13(e), 4.13(f), foram processadas pelos algoritmos de binarização de cada modelo e obteve-se os seguintes resultados, visto nas Figuras 4.14, 4.15, 4.16, 4.17, 4.18, 4.19.

Figura 4.14: Aplicação dos métodos estudados à imagem NAB01

(a) Original **(b)** *Niblack* **(c)** Ad. Média

(d) Ad. Gaussiana **(e)** Sauvola T= 166 **(f)** Otsu T= 139 **(g)** Proposto T= 86



(Fonte: Autor da Figura)

Dentre todas as Figuras vistas, na Figura 4.14, a Figura que apresentou os resultados mais eficientes em comparação com os outros modelos de binarização foi a Figura 4.14(g). A respectiva Figura 4.14(g), também, obteve um resultado capaz de filtrar os ruídos da frente e do verso da carta. Na imagem da Figura 4.14, os resultados, que mais se aproximam é o ad. média e o ad. Gaussiana, que por serem métodos de binarização local, conseguiram obter um resultado muito parecido com o modelo proposto estatístico pelo método global. Na Figura 4.15, a binarização, novamente, conseguiu obter um bom resultado ao eliminar o amarelado do papel e a retirada dos ruídos. Destaque para o bom resultado do modelo proposto. Na imagem, da Figura 4.15, os modelos, que mais se aproximaram do modelo proposto foram o Sauvola e Otsu. Os ad. pela média e pela Gaussiana não conseguiram obter um resultado satisfatório comparado com os outros modelos.

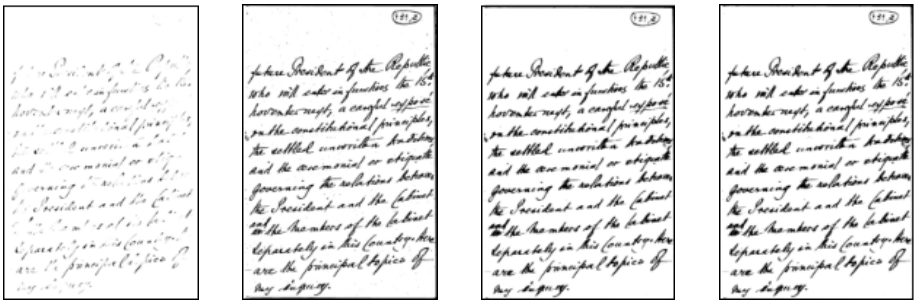
A Figura 4.16 obteve uma binarização satisfatória e assim como a Figura 4.15, a metodologia proposta resultou num resultado muito próximo ao modelo automático de Otsu.

Na Figura 4.17, a metodologia proposta não obteve o melhor resultado, mos-

Figura 4.15: Aplicação dos métodos estudados à imagem PRES01

(a) Original **(b)** *Niblack* **(c)** Ad. Média

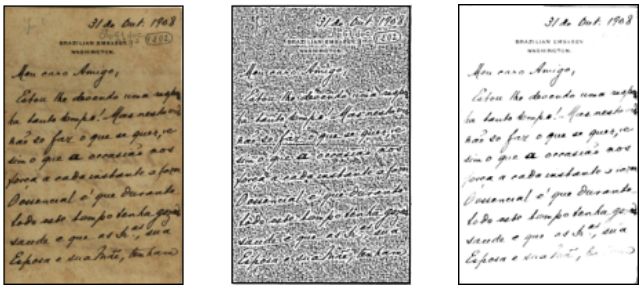
(d) Ad. Gaussiana **(e)** Sauvola T= 101 **(f)** Otsu T= 84 **(g)** Proposto T= 86



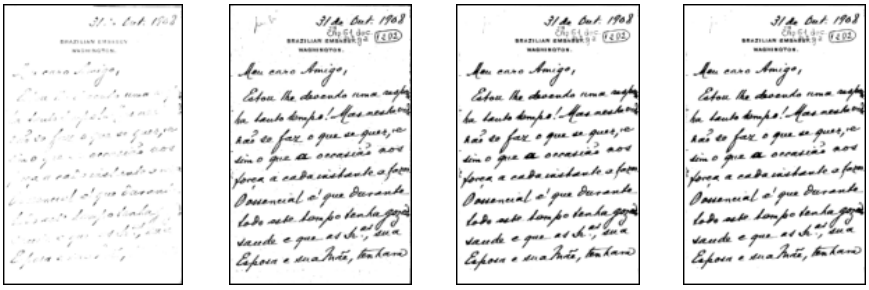
(Fonte: Autor da Figura)

Figura 4.16: Aplicação dos métodos estudados à imagem MCA01

(a) Original **(b)** *Niblack* **(c)** Ad. Média



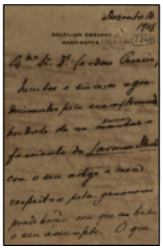
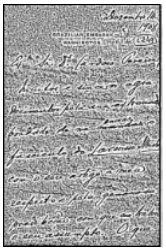
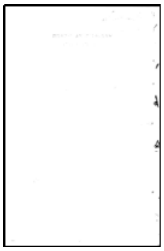
(d) Ad. Gaussiana **(e)** Sauvola T= 102 **(f)** Otsu T= 84 **(g)** Proposto T= 86



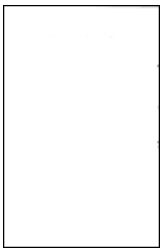
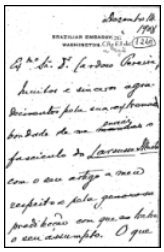
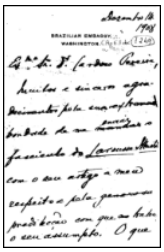
(Fonte: Autor da Figura)

Figura 4.17: Aplicação dos métodos estudados à imagem EBX01

(a) Original **(b)** *Niblack* **(c)** Ad. Média

(d) Ad. Gaussiana **(e)** Sauvola T= 75 **(f)** Otsu T= 66 **(g)** Proposto T= 86

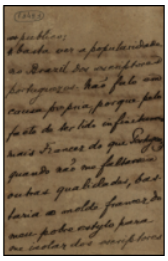
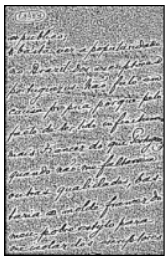
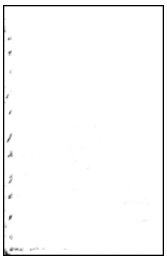





(Fonte: Autor da Figura)

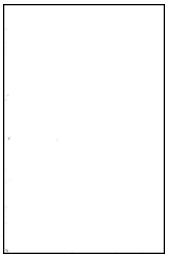
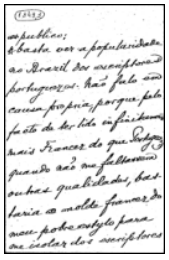
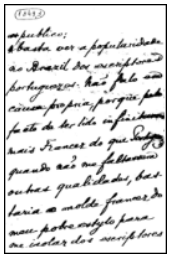
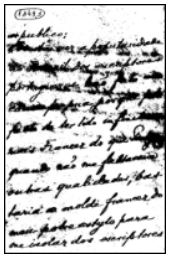
trando que o estudo para algumas amostras não foi eficiente para vencer o ruído do documento histórico em questão, assim como outros métodos clássicos que não prosperam. Mas, pelo menos o resultado proposto foi satisfatório na grande maioria da base de imagens. Nesta imagem da EBX01, os melhores resultados apresentados foram do modelo Sauvola conforme a Figura 4.17(e) e Otsu na Figura 4.17(f).

Figura 4.18: Aplicação dos métodos estudados à imagem ESC01

(a) Original **(b)** *Niblack* **(c)** Ad. Média

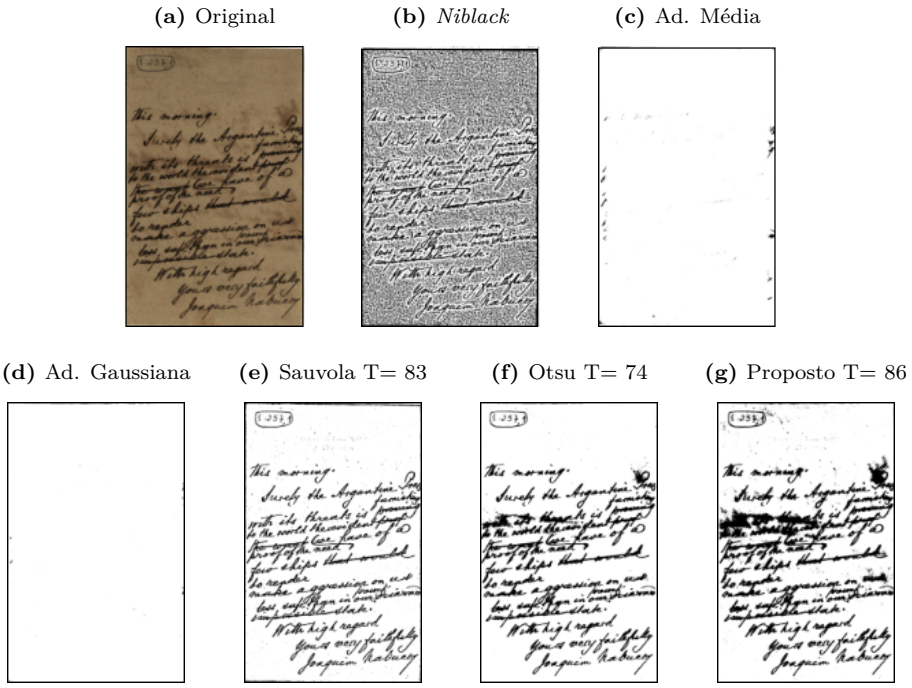




(d) Ad. Gaussiana **(e)** Sauvola T= 77 **(f)** Otsu T= 69 **(g)** Proposto T= 86

(Fonte: Autor da Figura)

Figura 4.19: Aplicação dos métodos estudados à imagem NAB02



(Fonte: Autor da Figura)

Nas Figuras 4.18 e 4.19, o método Sauvola é apresentado como o melhor resultado. No entanto, o resultado da metodologia proposta é muito próximo do modelo Sauvola, isto se deve a proximidade do *threshold* em ambos os modelos.

4.3 RESULTADOS ANALÍTICOS

Nesta seção, serão demonstrados analiticamente os resultados obtidos pelos modelos de binarização.

4.3.1 Relação Sinal-Ruído de Pico-PSNR

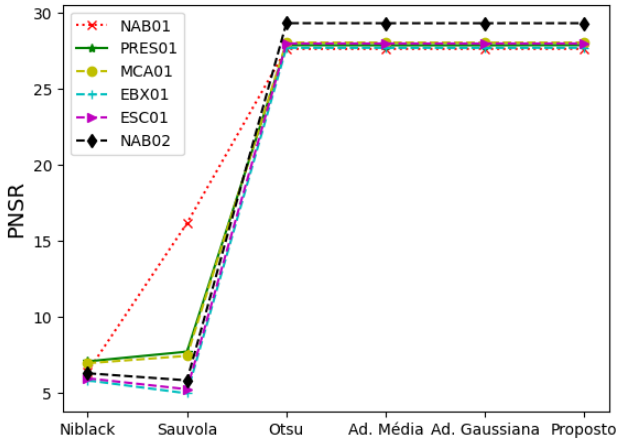
Com o intuito de comparar duas imagens, a original e as obtidas pelos modelos de binarização, foi calculado o PSNR para avaliar a qualidade de cada modelo, que pode ser visto na tabela 4.4, os resultados das amostras dessas imagens (NAB01, PRES01, MCA01, EBX01, ESC01, NAB02).

Tabela 4.4: Resultado da relação PSNR comparado com a imagem original

Modelos	NAB01	PRES01	MCA01	EBX01	ESC01	NAB02
Niblack	6,42078	7,094469	6,973339	5,844765	5,981739	6,313095
Sauvola	16,184244	7,734849	7,457574	4,99276	5,26205	5,846578
Otsu	27,641247	27,88295	28,054137	27,70977	27,99394	29,329716
Ad. Média	27,641372	27,876458	28,052898	27,708637	27,99309	29,326825
Ad. Gaussiana	27,641847	27,876519	28,051999	27,708529	27,993004	29,326543
Proposto	27,642162	27,887492	28,054526	27,712814	27,993933	29,32772

Embora o PSNR possa ser representado em uma faixa de 0 a 100 dB ou mais, em imagens e vídeos os valores úteis e típicos estão entre 20 a 50 dB, e a melhor forma de interpretar é entender, que quanto maior for o valor do PSNR, melhor será o modelo empregado, ou seja, melhor será o processo de binarização. Como visto na Tabela 4.4, os melhores resultados são Otsu, Adptativo pela média, adaptativo pela Gaussiana e o modelo proposto do *threshold*, mas, por possuírem valores muito próximos é difícil identificar qual o melhor resultado. Como mostra o gráfico não linear da Figura 4.20.

Figura 4.20: Comparação do PSNR da imagem original com os modelos de Binarização.



(Fonte: Autor da Figura)

Ao buscar uma forma de identificar o melhor resultado, foi proposto um padrão de normalidade, o qual, subtrai o valor mais baixo de cada imagem e multiplica-se

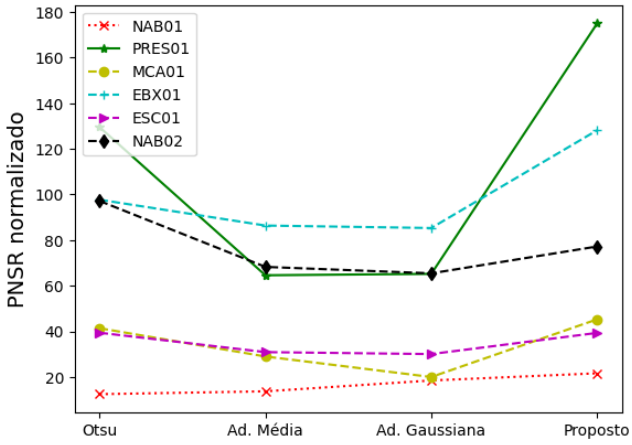
por 1.000, para o papel e para a escrita. Para exemplificar, de acordo com a tabela 4.4 pode-se observar, que os modelos da imagem NAB02 que obtiveram 29,32 foi o modelo Ad. Gaussiana, assim este resultado foi subtraído dos outros da imagem NAB02 e multiplicado por 1.000, para normalizar os dados. Dessa forma, foi obtido uma resolução melhor do processo, conforme a Tabela 4.5. Na tabela 4.5, pode-se observar, que os melhores resultados se concentram no modelo do *threshold*, tendo apenas uma imagem, que evidencia o modelo de Otsu como o melhor, a NAB02, mas por ela ter um PSNR muito próximo ao do modelo proposto, a qualidade dos resultados se assemelham nesses dois modelos.

Tabela 4.5: Comparação da Normalização dos Resultados do PSNR da imagem original com modelos de Binarização.

Modelos	NAB01	PRES01	MCA01	EBX01	ESC01	NAB02
Otsu	12,4699	129,5	41,37	97,6999	39,4	97,16
Ad. Média	13,7199	64,5799	28,9799	86,37	30,9	68,2499
Ad. Gaussiana	18,4699	65,1899	19,9899	85,2899	30,04	65,43
Proposto	21,6199	174,92	45,2689	128,14	39,3299	77,1999

Na Figura 4.21, no gráfico não linear, pode-se observar que, em 4 das 6 imagens, o melhor modelo é o proposto estatisticamente, porém o modelo Otsu supera o proposto, em apenas duas imagens (ESC01 e a NAB02).

Figura 4.21: Normalização da comparação do PSNR da imagem original com modelos de Binarização.



(Fonte: Autor da Figura)

4.3.2 Mapeamento de *Pixels* nas Imagens Binarizadas

Este método mapeia todos os *pixels* de uma imagem binarizada pelo modelo proposto e pelos métodos de binarização clássicos comparando-os, com o objetivo de buscar semelhanças e demonstrar quais modelos se aproximam do modelo proposto. Os resultados desta comparação de *pixels* são apresentados na Tabela 4.6.

Tabela 4.6: Mapeamento de Pontos Semelhantes nas Imagens Binarizadas em Comparação ao Modelo Proposto.

Modelos	NAB01	PRES01	MCA01	EBX01	ESC01	NAB02
Proposto	1.537.320	1.272.024	1.272.024	1.253.608	1.253.608	1.242.440
Niblack	1.085.503	972.740	926.535	868.082	893.497	893.723
Sauvola	1.498.641	1.248.395	1.249.037	1.167.606	1.175.619	1.182.698
Otsu	1.498.810	1.266.364	1.265.685	1.195.424	1.212.453	1.227.074
Ad. Média	1.521.139	1.199.068	1.231.779	1.087.268	1.072.010	1.067.269
Ad. Gaussiana	1.527.190	1.164.056	1.198.099	1.083.331	1.067.085	1.059.877

Na Tabela 4.6, observou-se os modelos, que mais se assemelham ao modelo proposto são os modelos Otsu e Sauvola, visto que, apresentam uma quantidade maior de *pixels* coincidentes, ao comparar com a quantidades de pontos do modelo proposto. Para uma melhor visualização, calculou-se uma porcentagem de semelhança, dividindo os valores de semelhança pelo total de *pixels* analisados e os resultados obtidos podem ser vistos na Tabela 4.7.

Tabela 4.7: Comparação Percentual do Nível de Semelhança com Base no Modelo Proposto.

Modelos	NAB01	PRES01	MCA01	EBX01	ESC01	NAB02
Proposto	100%	100%	100%	100%	100%	100%
Niblack	70,61%	76,47%	72,84%	69,25%	71,27%	71,93%
Sauvola	97,48%	98,14%	98,19%	93,14%	93,78%	95,19%
Otsu	97,49%	99,55%	99,50%	95,36%	96,72%	98,76%
Ad. Média	98,95%	94,26%	96,84%	86,73%	85,51%	85,90%
Ad. Gaussiana	99,34%	91,51%	94,18%	86,42%	85,12%	85,31%

Na Tabela 4.7, detectou-se incompatibilidade nas três últimas imagens (EBX01, ESC01, NAB02), mostrando que os modelos adaptativos pela média e pela Gaussiana se distanciam um pouco dos 90% de semelhança. Observou-se, também, que o modelo *Niblack* não apresenta um resultado próximo ao modelo baseado no *threshold*, indicando que a binarização desse modelo adaptativo não atende às expectativas, deixando o texto ilegível e com rasuras. Portanto, o modelo baseado na busca do *threshold* estatístico é o melhor de todos os modelos apresentados, seguido pelo Otsu como uma solução alternativa.

Capítulo 5

CONSIDERAÇÕES FINAIS

5.1 CONCLUSÃO

Este trabalho conclui que a metodologia proposta, por meio da aplicação de técnicas e algoritmos às amostras, obteve resultados satisfatórios na maioria das binarizações das imagens analisadas. Essa eficácia é atribuída ao tratamento adequado da base de dados, que envolve a análise e o ajuste dos dados das amostras do papel e da escrita, garantindo assim significância estatística.

Inicialmente, foi realizado um processo de amostragem na base das imagens com o objetivo de extrair características da textura do papel e da escrita. Este processo resultou em atributos, tais como, média, assimetria, curtose e outros. Assim, após experimentos empíricos, constatou-se que a média dos tons de cinza era a característica mais apropriada, pois por ser uma grandeza de primeira ordem, a variação desta grandeza provocava mudanças significativas no processo de binarização.

Com as amostras dos tons de cinza do papel e da escrita, realizou-se um teste de normalidade que revelou que ambas as amostras não apresentavam distribuições gaussianas. Então, para garantir a significância estatística dos dados, foi aplicada uma técnica chamada de *bootstrapping*, que gerou amostras mantendo o padrão estatístico original. Em seguida, foi refeito o teste de normalidade, que resultou em amostras normais para o papel e a escrita. Para assegurar a robustez e confiabilidade dos resultados, foi realizada uma análise de variância em cada amostra, e revelou a

ausência de diferenças significativas nos dados da amostra do papel e da amostra da escrita. Indicando que as médias dos tons de cinza podem ser atribuídas ao acaso, e não há diferença real entre os grupos da amostra do papel. Resultado semelhante é obtido em relação a amostra da escrita.

Os testes realizados serviram como base para o cálculo do *threshold* estatístico, fundamentado na textura dos documentos, garantindo uma binarização eficaz para a maior parte da população das imagens analisadas. Apesar desses resultados satisfatórios, houve casos em que os resultados não foram significativos. Então, para obter uma faixa de *threshold* confiável, realizou-se uma variação percentual do valor padrão, tanto para mais, quanto para menos. Ao aumentar o respectivo valor, permitiu-se uma deterioração da imagem, aumentando o ruído e ao reduzi-lo pode-se perder principalmente informações da escrita.

Como forma de validar os resultados, além das imagens apresentadas, calculou-se o PSNR para definir o melhor modelo de binarização de forma analítica. Quanto maior o valor do PSNR, maior a eficácia da binarização. Os resultados mostraram que modelos como Otsu e o modelo proposto se destacaram na maioria das binarizações. Outro método de avaliação, consistiu em identificar qual modelo se aproximava mais do modelo proposto pelo mapeamento de *pixels*, revelando que Otsu, um modelo consagrado nos modelos clássicos de binarização, foi o que mais se aproximou do modelo proposto. Dessa forma, a proposta se torna um estudo válido para o processamento digital de imagens, pois o modelo de binarização atende à necessidade de inferir documentos binarizados, filtrando ruídos indesejados, como os provenientes da frente e do verso do documento, bem como os efeitos do envelhecimento do papel. Este processo foi realizado utilizando um *threshold* baseado em estatística, tornando a binarização dinâmica. Isso proporciona uma melhor compreensão das imagens e, consequentemente, um impacto positivo no armazenamento de informações históricas.

5.2 TRABALHOS FUTUROS

Para trabalhos futuros, seria interessante digitalizar materiais históricos, como cartas e textos utilizando diferentes condições de iluminação. Esta abordagem ajudaria a captar uma gama mais ampla de variações de dados, permitindo uma binrização mais rica e precisa. Outra sugestão é trabalhar com uma base de dados maior, pois, dessa forma, melhorará o desempenho do modelo, reduzindo a chance de *overfitting*, que é quando o modelo se ajusta muito bem aos dados de treinamento, mas não se sai bem com dados novos. Finalmente, utilizar técnicas e ferramentas de *data science* (área interdisciplinar que utiliza métodos científicos, processos, algoritmos e sistemas para extrair conhecimentos de dados) para buscar o aprimoramento do trabalho proposto.

Referências

Bertholdo, F. A. R. (2007). Técnicas de limiarizaçao para melhorar a qualidade visual de documentos históricos.

Broch, S. C. e Ferreira, D. F. (2012). Estudo do poder de teste de hipoteses para a variˆancia populacional. *Universidade de São Paulo*.

Crosta, A. P. (1999). *Processamento digital de imagens de sensoriamento remoto*. UNICAMP/Instituto de Geociências.

da Silva Filho, A. S. (2010). Inferência em amostras pequenas: método bootstrap. *Revista de Ciências exatas e tecnologia*, 5(5):115–126.

Gonzalez, R. C. e Woods, R. E. (2010). *Processamento digital de imagem*.

Hirakata, V. N., Mancuso, A. C. B., e de Jezus Castro, S. M. (2019). Teste de hipóteses: perguntas que você sempre quis fazer, mas nunca teve coragem. *Clinical and Biomedical Research*, 39(2).

Ikeda, D. T. (2011). Seleção não-supervisionada de métodos de binarização para documentos históricos. *Departamento de Ciência da Computação, Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo*.

Ito, R. H., Kim, H. Y., e Salcedo, W. J. (2009). Classificação de texturas invariante a rotação usando matriz de co-ocorrência. In *8th International Information and Telecommunication Technologies Symposium*, páginas 1–6.

Lopes, F. M. (2003). Um modelo perceptivo de limiarização de imagens digitais. *Universidade Federal do Paraná*, páginas 7,8,9.

Marques Filho, O. e Neto, H. V. (1999). *Processamento digital de imagens*. Brasport.

Meneses, P. R. e Almeida, T. d. (2012). Introdução ao processamento de imagens de sensoriamento remoto. *Universidade de Brasília, Brasília*, páginas 183,184.

Nunes, F. L. (2006). Introdução ao processamento de imagens médicas para auxílio a diagnóstico–uma visão prática. *Livro das Jornadas de Atualizações em Informática*, páginas 56,$_9$2, 98.

Oliveira, M. M. et al. (2019). Avaliação de ruído em perfil de imagens mamográficas.

Paese, C., Caten, C. t., e Ribeiro, J. L. D. (2001). Aplicação da análise de variância na implantação do cep. *Production*, 11:17–26.

Pereira, M. V. L. (2006). Reconhecimento de caracteres estampados em tarugos de aço.

Pino, F. A. (2014). A questão da não normalidade: Uma revisão. *Revista de economia agrícola*, 61(2):17–33.

Ribas, F. B. et al. (2013). Análise de técnicas de limiarização em programas de análise de cartões hidrossensíveis.

Schwartz, W. R. e Pedrini, H. (2003). Método para classificação de imagens baseada em matrizes de coocorrência utilizando características de textura. *III Colóquio Brasileiro de Ciências Geodésicas, CuritibaPR, Brasil*, 1.

Seixas, F. L., Martins, A., Stilben, A. R., Madeira, D., Assumpção, R., Mansur, S., Victer, S. M., Mendes, V. B., e Conci, A. (2008). Avaliaçao dos métodos para a segmentaçao automática dos tecidos do encéfalo em ressonância magnética. *Simpósio de Pesquisa Operacional e Logıstica da Marinha SPOLM*.

Torok, L. (2016). Método de otsu. *Instituto de Computação (UFF)*.

Victer, S. M. e Silva, B. S. (2019). Estudo comparativo de técnicas de limiarização de histogramas em imagens de nódulos de mamografia. *Revista Mundi Engenharia, Tecnologia e Gestão (ISSN: 2525-4782)*, 4(3).

Anexo 1

Expressões Estatísticas Baseadas No Histograma

Nas fórmulas abaixo, p(i) é um vetor de histograma normalizado, cujas as entradas são divididas pelo número total de *pixels* em ROI, i=1,2,.., N_g. Sendo N_g níveis de intensidade.

Média, valor que descreve a concentração de dados de uma distribuição, sendo expressa pela seguinte equação:

$$\mu = \sum_{i=1}^{N_g} ip(i) \quad . \tag{1.1}$$

Variância, é uma medida de dispersão estatística, que indica a diferença em relação a sua média, expresso pela seguinte equação:

$$\sigma^2 = \sum_{i=1}^{N_g} (i-\mu)^2 p(i) \quad . \tag{1.2}$$

Skewness, é uma medida de falta de simetria de uma determinada distribuição de frequência, expresso pela seguinte equação:

$$\mu_3 = \sigma^{-3} \sum_{i=1}^{N_g} (i-\mu)^3 p(i) \quad . \tag{1.3}$$

Curtose, termo usado na teoria da probabilidade e estatística, para descrever o

pico de uma distribuição de frequência,expresso pela seguinte equação:

$$\mu_4 = \sigma^{-4} \sum_{i=1}^{N_g} (i - \mu)^4 p(i) - 3 \quad . \tag{1.4}$$

Anexo 2

Expressões Estatísticas Baseadas Na Matriz de Co-ocorrência

O histograma de segunda ordem é definido como a matriz de co-ocorrência $h_{d\theta}(i,j)$. Quando dividido pelo total número de *pixels* vizinhos $R(d,\theta)$ em ROI, esta matriz se torna a estimativa da probabilidade conjunta, $p_{d\theta}(i,j)$, de dois *pixels*, a uma distância d ao longo de uma determinada direção θ tendo valores particulares (co-ocorrentes) i e j. Formalmente, dada a imagem $f(x,y)$ com um conjunto de N_g níveis de intensidade discretos, a matriz $h_{d\theta}(i,j)$ é definido tal que sua (i,j)ésima entrada seja igual ao número de vezes que

$$\begin{aligned} f(x_1,y_1) = i \;\; and \;\; f(x_2,y_2) = j \;\; . \\ Where \;\; (x_2,y_2) = (x_1,y_1) + (dcos\theta, dsen\theta) \;\; . \end{aligned} \tag{2.1}$$

Isso produz uma matriz quadrada de dimensão igual ao número de níveis de intensidade na imagem, para cada distância d e orientação θ;. No MaZda®, as distâncias d = 1, 2, 3, 4 e 5 *pixels* com ângulos $\theta = 0°$, 45°, 90° e 135° são considerados. Redução do número de níveis de intensidade (por quantização para menos níveis de intensidade) ajuda a aumentar a velocidade da computação, com alguma perda de informação textural.

Os parâmetros derivados da matriz de co-ocorrência calculados pelo MaZda® são definidos pelas equações, onde μ_x, μ_y e σ_x, σ_y denotam a média e os desvios padrão das somas das linhas e colunas de a matriz de co-ocorrência, respectivamente (relacionada às distribuições marginais $p_x(i)$ e $p_y(j)$).

Angular second moment:

$$AngScMom = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)^2 \quad . \tag{2.2}$$

Contrast:

$$Contrast = \sum_{n=0}^{N_g-1} n^2 \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j) \quad . \qquad |i-j| = n \tag{2.3}$$

Correlation:

$$Correlat = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} ijp(i,j) - \mu_x\mu_y}{\rho_x\rho_y} \quad . \tag{2.4}$$

Sum of squares:

$$SumOfSqs = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i-\mu_x)^2 p(i,j) \quad . \tag{2.5}$$

Inverse difference moment:

$$InvDfMom = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\frac{1}{1+(i-j)^2}p(i,j) \quad . \tag{2.6}$$

Sum average:

$$SumAverg = \sum_{i=1}^{2N_g} ip_{x+y}(i) \quad , \qquad Where \quad p_{x+y}(k) = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j) \quad k-2,3...,2N_g \qquad i+j=k \tag{2.7}$$

Sum variance:

$$SumVarnc = \sum_{i=1}^{2N_g} (i - SumAverg)^2 p_{x+y}(i) \tag{2.8}$$

Sum entropy:

$$SumEntrp = -\sum_{i=1}^{2N_g} p_{x+y}(i) log(p_{x+y}(i)) \tag{2.9}$$

Entropy:

$$Entropy = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) log(p(i,j)) \tag{2.10}$$

Difference variance:

$$DifVarnc = \sum_{i=0}^{N_g-1} \sum_{j=1}^{N_g} (i - \mu_{x-y})^2 p_{x-y}(i) \quad , \tag{2.11}$$

Onde μ_{x+y} é um valor médio da distribuição de diferença

$$p_{x-y}(k) = \sum_{i=0}^{N_g} \sum_{j=1}^{N_g} p(i,j) \qquad k = 0, 1, ..., N_g - 1 \qquad |i - j| = k \tag{2.12}$$

Difference entropy:

$$DifEntrp = -\sum_{i=1}^{N_g-1} p_{x-y}(i) log(p_{x-y}(i)) \quad , \tag{2.13}$$

Anexo 3

Algoritmos Que Executam Binarizações Globais e Adaptativas

```
import glob
folder= 'base/*'
image_files_list = glob.glob(folder)
import cv2
import pandas as pd
import cv2 as cv
import numpy as np
from matplotlib import pyplot as plt
def showSingleImage(img,title,size):
    fig, axis = plt.subplots(figsize= size)
    axis.imshow(img, 'gray')
    axis.set_title(title, fontdict={'fontsize':22, 'fontweight': '
   medium'})
    plt.show()
img_filename= image_files_list[1]
img= cv2.imread(img_filename)#img_filename)
img = cv2.cvtColor(img, cv2.COLOR_BGR2GRAY)
thresh,img_thresh1B= cv2.threshold(img,86,255,cv2.THRESH_BINARY)
showSingleImage(img_thresh1B, "Binariza  o por THRESHOLD", (5,5))
```

Figura 3.1: Quadro 5 - Algoritmo de Binarização pelo Método Proposto

```
import glob
folder= 'base/*'
image_files_list = glob.glob(folder)
import cv2
import pandas as pd
import cv2 as cv
import numpy as np
from matplotlib import pyplot as plt
def showSingleImage(img,title,size):
    fig, axis = plt.subplots(figsize= size)
    axis.imshow(img, 'gray')
    axis.set_title(title, fontdict={'fontsize':22, 'fontweight': '
   medium'})
    plt.show()
import cv2 as cv
block_size=99 # pedacos so pode ser numeros impares
C=86 # quanto mais alto maior sera a diferenca entre escrita e
   papel(tela d fundo e objeto) so pega no adptativo e no gaussiano
img_filename= image_files_list[5]#0 a 39 ##18,20
img= cv2.imread(img_filename)#img_filename)
img = cv2.cvtColor(img, cv2.COLOR_BGR2GRAY)
imgAdapGauss = cv2.adaptiveThreshold(img, 255, cv2.
   ADAPTIVE_THRESH_GAUSSIAN_C, cv2.THRESH_BINARY,block_size,C)
showSingleImage(imgAdapGauss, "Binarizacao Adptativa pela Gaussiana
   ", (5,5))
```

Figura 3.2: Quadro 6 - Algoritmo da Binarização pela Gaussiana

```
img= cv2.imread(img_filename)#img_filename)
img = cv2.cvtColor(img, cv2.COLOR_BGR2GRAY)
showSingleImage(img, "00",(5,5))
imgOriginal = img
limiar, otsu= cv2.threshold(imgOriginal, 0, 255, cv2.THRESH_OTSU)
showSingleImage(otsu, "Binarizacao Otsu =" +str(limiar),(5,5))
```

Figura 3.3: Quadro 7 - Algoritmo de Binarização Otsu

```
import glob
folder= 'base/*'
image_files_list = glob.glob(folder)
import cv2
import pandas as pd
import cv2 as cv
import numpy as np
from matplotlib import pyplot as plt
def showSingleImage(img,title,size):
    fig, axis = plt.subplots(figsize= size)
    axis.imshow(img, 'gray')
    axis.set_title(title, fontdict={'fontsize':22, 'fontweight': '
    medium'})
    plt.show()
block_size=99 # peda os so pode ser n meros  mpares
C=86 # quanto mais alto maior ser  a diferen a entre escrita e
    papel(tela d fundo e objeto) so pega no adptativo e no gaussiano
img_filename= image_files_list[5]#0 a 39 ##18,20
img= cv2.imread(img_filename)#img_filename)
img = cv2.cvtColor(img, cv2.COLOR_BGR2GRAY)
imgAdapMean = cv2.adaptiveThreshold(img, 255, cv2.
    ADAPTIVE_THRESH_MEAN_C, cv2.THRESH_BINARY,block_size,C)
showSingleImage(imgAdapMean, "Binarizacao Adptativa pela media",
    (5,5))
```

Figura 3.4: Quadro 8 - Algoritmo da Binarização pela Média

```
import glob
folder= 'base/*'
image_files_list = glob.glob(folder)
import cv2
import pandas as pd
import cv2 as cv
import numpy as np
from matplotlib import pyplot as plt
img_filename= image_files_list[1]
img= cv2.imread(img_filename)#img_filename)
img = cv2.cvtColor(img, cv2.COLOR_BGR2GRAY)
# Importing necessary libraries
from skimage import data
from skimage import filters
from skimage.color import rgb2gray
import matplotlib.pyplot as plt
NI=img
# Computing Ni black's local pixel
# threshold values for every pixel
threshold = filters.threshold_niblack(NI)
# Computing binarized values using the obtained
# threshold
binarized_niblack = (NI > threshold)*1
#plt.subplot(2,2,2)
plt.title("Niblack")
plt.imshow(binarized_niblack, cmap = "gray")
```

Figura 3.5: Quadro 9 - Algoritmo de Binarização *Niblack*

```
import glob
folder= 'base/*'
image_files_list = glob.glob(folder)
import cv2
import pandas as pd
import cv2 as cv
import numpy as np
from matplotlib import pyplot as plt
img_filename= image_files_list[1]
img= cv2.imread(img_filename)#img_filename)
img = cv2.cvtColor(img, cv2.COLOR_BGR2GRAY)
# Importing necessary libraries
from skimage import data
from skimage import filters
from skimage.color import rgb2gray
import matplotlib.pyplot as plt
SA=img
threshold = filters.threshold_sauvola(SA)
#plt.subplot(2,2,3)
plt.title("Sauvola")
plt.imshow(threshold, cmap = "gray")
# Computing Sauvola's local pixel
# threshold values for every pixel - Binarized
binarized_sauvola = (SA > threshold)*1
plt.title("Sauvola Thresholding - Converting to 0's and 1's")
# Displaying the binarized image
plt.imshow(binarized_sauvola, cmap = "gray")
```

Figura 3.6: Quadro 10 - Algoritmo de Binarização Sauvola